

Rockefeller Series in Science and Technology 4

Alessandro N. Vargas *Editor*

Jyoti Verma
Arpita Srivastava
Subir Verma

Diatom Indexing with Diathor in RStudio

A practical guide

ISSN 3067-0667



ROCKEFELLER
NEW YORK, NY



Rockefeller Series in Science and Technology

Open Access • Book Series

Volume 4

Series Editor

Alessandro N. Vargas  • UTFPR Brazil • UC Berkeley USA

For further volumes:

<https://www.rockefellerpub.com/series>

The Rockefeller Series in Science and Technology (ISSN 3067-0667) publishes research monographs, edited works, collections of papers, reviews of previously published studies, and advanced textbooks covering all relevant subjects related to science and technology.

Monographs of this series can contain the following contributions:

- Book chapters that revisit and expand the interpretation of previously published works;
- Timely discussions of a relevant topic;
- Literature review with insights and novel interpretations;
- A carefully designed study that sparks interest in the research community;
- Research tailored for graduate students and professionals;
- Technical reports;
- Other topics related to the book series.

The monographs published in the Rockefeller Series on Science and Technology attract the interest of researchers, students, and professionals.

Rockefeller Publishing Co. welcomes book ideas. Potential authors who wish to submit a book proposal should contact the Editors at <https://rockefellerpub.com/books>.

Diatom Indexing with Diathor in RStudio: A practical guide

Jyoti Verma  • University of Allahabad, India

Arpita Srivastava  • University of Allahabad, India

Subir Verma • Amazon, India



Jyoti Verma 

Arpita Srivastava 

Aquatic Ecology Lab, Department of Zoology, CMP Degree College, University of Allahabad, Prayagraj, India

Subir Verma

Amazon, Bangalore-560035, India

ISSN 3067-0667

Rockefeller Series in Science and Technology

ISBN 979-8-9928364-3-1 (eBook)

DOI: <https://doi.org/10.63408/979-8-9928364-3-1>

© The Author(s) 2026. This book is an open access publication.

Notice. This book is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0). This license allows authors to publish their work openly, enabling others to use, share, adapt, distribute, and reproduce the content in any medium or format, as long as appropriate credit is given to the original author(s) and source.

The publisher, authors, and editors believe that the information presented in this book is accurate as of the publication date. However, they do not offer any guarantee, express or implied, regarding the accuracy or completeness of the material contained herein, nor do they assume responsibility for any errors or omissions. The application of the information in this book should be taken at the reader's discretion.

The publisher remains neutral about jurisdictional claims in published maps and institutional affiliations.

This open-access book is published and distributed online by Rockefeller Publishing Co.

The typesetting and image files were prepared by the authors.

Rockefeller Publishing Co., 45 Rockefeller Plaza 20th Floor, New York, NY 10111 USA.

Diatom Indexing with Diathor in RStudio: A practical guide

Jyoti Verma • Arpita Srivastava • Subir Verma



Abstract: Aquatic biomonitoring has historically relied on the meticulous, yet profoundly labour-intensive, manual taxonomic identification and mathematical calculation of diatom communities. As environmental datasets grow in scale and complexity, the discipline faces a critical demand for reproducible, high-throughput analytical frameworks. *Diatom Indexing with DiaThor in RStudio: A Practical Guide* provides a methodological synthesis of classical freshwater algal ecology and modern computational data science. By leveraging the open-source DiaThor R package, this text equips aquatic ecologists, taxonomists, and environmental managers with the programmatic rigor required to transition from raw microscopy counts to sophisticated ecological insights. The book systematically deconstructs the computational pipeline for diatom bioassessment. It addresses foundational data tidying principles, rigorous formatting protocols, and the algorithmic computation of globally recognized biotic indices (e.g., IPS, TDI, SLA), alongside functional and morphological trait classifications such as Van Dam ecological preferences and habitat guilds. Beyond quantitative metrics, the text places a heavy emphasis on graphical communication, guiding users through advanced visualization techniques. By detailing the use of libraries like ggplot2 and vegan for multivariate ordinations, spatial mapping, and trait-based clustering, the authors ensure that empirical data are translated into actionable environmental narratives. Ultimately, this guide demystifies the RStudio environment for non-programmers, fostering a vital paradigm of transparent, reproducible research in aquatic ecosystem monitoring.

Keywords: Aquatic biomonitoring • Diatom Indexing • DiaThor • RStudio • Biotic indices • Computational data science



DESCRIPTION



Diatom Indexing with DiaThor in RStudio: Code, Plot, Assess- A Practical Guide” is written for those who live and breathe ecology but want to harness the full power of data science without losing the soul of fieldwork. This isn't just another R manual ; it's a biologist's field guide to the digital world of diatoms. “

Built around the DiaThor package in R, this book helps you transform your diatom data into ecological wisdom from calculating classic biotic indices like IPS, TDI, and SLA, to generating elegant, publication-ready plots, heatmaps, radar charts, and trait visualizations. Every section is packed with copy-and-paste-ready R code, written in plain language, so you can go straight from dataset to interpretation without hours of syntax frustration.

Think of it as your lab partner in print, guiding you through the steps of loading, cleaning, analyzing, and visualizing ecological data, while also helping you understand what each output means in environmental terms. Whether you're assessing pollution, studying trophic gradients, or exploring diatom community shifts, this book shows how to use RStudio as an ecological microscope, magnifying patterns invisible to the naked eye.

From undergraduate learners to professional researchers, anyone working with freshwater monitoring or algal ecology will find this book an essential shortcut to clarity. It merges the art of environmental observation with the precision of computational analysis, empowering biologists to think, code, and visualize like data scientists, without losing touch with the rhythm of rivers and lakes.

In short, this guide doesn't just teach you R. it helps you speak ecology in the language of code. Let's understand in short, what's the story behind each chapters.

Chapter 1- Understanding Diatoms in Ecology: The Glass-Walled Guardians

This opening chapter introduces diatoms — the microscopic, silica-shelled architects of aquatic ecosystems. It explores their evolutionary beauty, ecological roles, and how these single-celled organisms act as environmental sentinels. It lays the philosophical and biological groundwork for why diatoms are central to water quality assessment and biomonitoring

Chapter 2- From Microscopes to Machine: The Need for Automation

Here, the narrative shifts from the classical to the computational. The chapter traces how traditional diatom identification under a microscope, though precise, is laborious — and how digital tools, data science, and automation revolutionize diatom research. It makes the case for adopting computational packages like DiaThor for efficient, reproducible ecological analysis.



Chapter 3- R Studio and DiaThor

This section familiarizes readers with RStudio, the open-source platform that powers most ecological analytics, and introduces DiaThor, the specialized R package for diatom-based water quality assessment. It explains the philosophy of reproducible coding, environment setup, and why R is the language of modern ecological science.

Chapter 4- Setting the stage: R Studio and DiaThor

This chapter introduces the installation and foundational use of RStudio and DiaThor for efficient diatom-based ecological analysis and water quality assessment.



Chapter 5- Data Dictionary for DiaThor Input Files & Troubleshooting Guide for DiaThor

An essential reference chapter, it dissects the structure and meaning of every input column in DiaThor's datasets — from sample IDs to taxonomic abundance. It doubles as a troubleshooting guide, offering solutions for common user errors like unrecognized taxa, missing indices, or database mismatches.

Chapter 6- How DiaThor Works?

This chapter unveils the internal workflow of DiaThor: how raw diatom counts translate into ecological indices. It breaks down the algorithmic logic, database dependencies, and the statistical backbone behind the computation of indices such as IPS, TDI, and others. Think of it as looking under the hood of the machine.

Chapter 7- Voice of the Water: Through the Lens of Diatom Indices

Here, the poetry of ecology meets the precision of data. The chapter decodes what diatom indices mean — how they give voice to the health, pollution level, and trophic state of water bodies. It interprets ecological indices not as numbers, but as narratives about environmental integrity.

Chapter 8- What is a Tidy Format?

Before data can sing, it must be tidy. This short yet crucial chapter explains the tidy data philosophy — how properly structured datasets make analysis in R more seamless. Using examples from diatom datasets, it illustrates reshaping, cleaning, and structuring for compatibility with packages like ggplot2 and dplyr.

Chapter 9- Essential R Packages for Ecological Visualization

A tour through the R ecosystem — ggplot2, tidyverse, plotly, ggraph, and more — showing how each contributes to ecological data visualization. The chapter highlights why visual literacy matters in environmental science and how to select the right tool for the right message.

Chapter 10- Data Dictionary for DiaThor Input Files & Troubleshooting Guide for DiaThor

An essential reference chapter, it dissects the structure and meaning of every input column in DiaThor's datasets — from sample IDs to taxonomic abundance. It doubles as a troubleshooting guide, offering solutions for common user errors like unrecognized taxa, missing indices, or database mismatches.

Chapter 11- A Methodological Start: DiaThor in Practice

The real hands-on section: how to run DiaThor step by step. It guides readers from loading data and calling core functions to interpreting the output files. Ideal for practitioners who prefer doing over reading, this chapter transforms theory into action.

Chapter 12- Quick Fix Steps for Commonly Encountered Errors

Because coding rarely goes smoothly, this section serves as a survival guide. It catalogues frequent user mistakes and provides crisp, actionable fixes — from handling NULL outputs to aligning species names with databases.

Chapter 13- Standard Image Size for Plot (in pixels)

A technical yet handy reference listing recommended resolutions and aspect ratios for publication-quality plots. It ensures visual consistency across manuscripts, posters, and digital dashboards.

Chapter 14- Drawing Plots from DiaThor Outputs

This chapter provides ready-to-run R scripts for drawing customized plots directly from DiaThor outputs. It emphasizes reproducibility and adaptability — users can modify code snippets to match their datasets and aesthetic preferences.

Chapter 15- One Index at a Time: Custom Analysis with DiaThor

Here, analysis becomes granular. Instead of running the full `DiaThor_all()` function, this chapter shows how to compute and visualize individual indices selectively — empowering researchers to design focused, hypothesis-driven ecological studies.

Chapter 16- Parameters for Interpreting the Result of Various Calculated Indices

A glossary and interpretation guide. It defines each ecological index and provides reference scales for pollution levels, saprobity, and trophic status. This transforms raw numbers into ecological meaning.

Chapter 17- The Ultimate Toolkit for Plotting Any Index

A consolidated collection of versatile R functions for plotting any ecological index — compatible with or without DiaThor. It's like a Swiss Army knife for visualization enthusiasts.

Chapter 18- Diatom Indices with DiaThor: Code, Plots, and Ecological Interpretation

The powerhouse chapter — a comprehensive walkthrough combining data input, computation, plotting, and interpretation. It shows how to tell ecological stories visually and numerically using DiaThor outputs.

Chapter 19- Cheat Code

A rapid-reference guide for power users. Lists concise, optimized R commands, shortcuts, and common patterns to automate frequent tasks. Think of it as the “speedrun” version of the previous chapters.

Chapter 20- Super Advanced CSV Plotter with Recommendations

Introduces a high-level R script capable of importing multiple CSV files (indices, guilds, traits) and generating composite visualizations. It demonstrates automation for batch plotting, ideal for large monitoring programs.

Chapter 21- How Useful is the DiaThor Package?



A reflective chapter evaluating DiaThor's scientific relevance. It balances its strengths — automation, reproducibility, transparency — against its limitations and explores its potential integration into environmental monitoring frameworks.

Chapter 22- Limitations of Using DiaThor and Its Plotting Functions

The closing critique. It discusses constraints such as limited index coverage, reliance on reference databases, and the need for taxonomic updates. The chapter concludes with a call for collaborative database enrichment and open-source ecological innovation.



INDEX

CONTENTS

DESCRIPTION	3
PREFACE	9
ABOUT THE AUTHORS	10
ACKNOWLEDGEMENT	11
INTRODUCTION	12
CHAPTER 1	14
UNDERSTANDING DIATOMS IN ECOLOGY: THE GLASS-WALLED GUARDIANS	14
CHAPTER 2	18
FROM MICROSCOPES TO MACHINE: THE NEED FOR AUTOMATION	18
CHAPTER 3	21
R STUDIO AND DIATHOR	21
CHAPTER 4	24
SETTING THE STAGE: R STUDIO AND DIATHOR	24
CHAPTER 5	28
DATA DICTIONARY FOR DIATHOR INPUT FILES & TROUBLESHOOTING GUIDE FOR DIATHOR	28
CHAPTER 6	31
HOW DIATHOR WORKS?	31
CHAPTER 7	34
VOICE OF THE WATER: THROUGH THE LENS OF DIATOM INDICES	34
CHAPTER 8	40
WHAT IS A TIDY FORMAT?	40
CHAPTER 9	42
ESSENTIAL R PACKAGES FOR ECOLOGICAL VISUALIZATION	42

CHAPTER 10	50
A METHODOLOGICAL START: DIATHOR IN PRACTICE	50
CHAPTER 11	74
PLOTTING AND VISUALIZATION	74
CHAPTER 12	80
QUICK FIX STEPS FOR COMMONLY ENCOUNTERED ERRORS	80
CHAPTER 13	84
STANDARD IMAGE SIZE FOR PLOT (IN PIXELS)	85
CHAPTER 14	88
DRAWING PLOTS FROM DIATHOR OUTPUTS	88
CHAPTER 15	94
ONE INDEX AT A TIME: CUSTOM ANALYSIS WITH DIATHOR	94
CHAPTER 16	129
PARAMETERS FOR INTERPRETING THE RESULT OF VARIOUS CALCULATED INDICES	129
CHAPTER 17	133
THE ULTIMATE TOOLKIT FOR PLOTTING ANY INDEX	133
CHAPTER 18	135
DIATOM INDICES WITH DIATHOR – CODE, PLOTS, AND ECOLOGICAL INTERPRETATION	135
CHAPTER 19	162
CHEAT CODE	162
CHAPTER 20	167
SUPER ADVANCED CSV PLOTTER WITH RECOMMENDATIONS	167
CHAPTER 21	181
HOW USEFUL IS DIATHOR PACKAGE?	181
CHAPTER 22	186
LIMITATIONS OF USING DIATHOR AND ITS PLOTTING FUNCTIONS	186

APPENDIX	190
REFERENCES	191
GLOSSARY	207
MESSAGE FROM THE AUTHOR	213

PREFACE

Rivers are more than flowing waters; they are living archives of time, memory, and change. Within them, diatoms reside; silent sentinels, fragile yet profound, reflecting the story of every ecological shift with remarkable clarity. To study them is to discover that even the smallest of beings carry within them the wisdom of an entire ecosystem.

*This book, *Indexing Diatom with DiaThor in R Studio*, was conceived with the vision of making diatom-based ecological analysis more approachable, reproducible, and meaningful for students, researchers, and educators. It integrates the biological significance of diatoms with the power of computational tools such as *DiaThor* and *R Studio*, presenting a bridge between ecological understanding and modern data science. The goal has been simple: to make the bridge between ecological knowledge and computational practice more accessible. Too often, ecological indices and their calculations are locked behind cryptic code or fragmented literature. Here, I have tried to walk the reader gently but firmly through that maze, chapter by chapter, showing not just the “how” but also the “why” behind each analysis.*

*This is not a book only about code, nor solely about ecology. It is about their union: the way statistical rigor meets biological intuition to generate insight. It is meant for students just dipping their toes into R, for scholars wrestling with *DiaThor* outputs, and for practitioners who need clear, interpretable tools to guide water management. If this book inspires even a few readers to see diatoms not as obscure data points but as storytellers of ecological resilience and fragility, it will have served its purpose.*

Our aim is not only to provide a practical guide to indices, coding, and visualization, but also to highlight the deeper message diatoms carry: that life, even in its smallest forms, can reflect resilience, balance, and change.

We hope this work serves as both a technical manual and a philosophical companion, inspiring readers to view ecological analysis not merely as science, but as a dialogue with nature.

ABOUT THE AUTHORS

Dr. Jyoti Verma is an Assistant Professor in the Department of Zoology at C.M.P. Degree College, University of Allahabad, Prayagraj. She has been actively engaged in research on diatoms and aquatic ecology, contributing to the understanding of how these microorganisms serve as bioindicators of environmental change. With years of teaching and research experience, Dr. Verma has guided several students in their academic journeys, combining scientific rigor with a passion for ecological conservation. Her expertise in diatom taxonomy and biomonitoring has been a strong foundation for this book.

Arpita Srivastava is a Research Scholar in the Department of Zoology at C.M.P. Degree College, Prayagraj., specializing in diatom-based water quality assessment. Her doctoral research focuses on pollution studies in the Yamuna River, with a special emphasis on trace and toxic metal contamination and biomonitoring indices. Her work bridges ecology, data science, and sustainable innovation. Her research explores the intricate relationships between aquatic ecosystems and human impact, using R programming, statistical modeling, and diatom-based bioassessment tools to decode environmental signals hidden in data. She has authored this book to integrate her hands-on experience with open-source computational tools such as DiaThor and R Studio. Arpita's work reflects her commitment to making ecological analysis more accessible for young researchers.

Subir Verma is a Senior Applied Scientist, currently working at Amazon and has co-authored this book. Driven by the art of turning research into reality. With a keen focus on natural language understanding, vector search, recommendation systems, and reinforcement learning, he builds intelligent systems that bridge theory and impact. Proficient in Python and grounded in modern MLOps practices—from Airflow and Docker to AWS, Azure, and FastAPI; Subir brings clarity and structure to the ever-evolving world of AI. Known for his pragmatic approach—start small, measure honestly, and refine relentlessly—he thrives in collaborative spaces, mentoring teams and designing experiments that move ideas forward. To him, data science isn't just about algorithms; it's about curiosity, precision, and the quiet pursuit of better answers.

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to our colleagues, mentors, and institutions for their support and encouragement. We also thank our families, who stood with patience and faith throughout this journey. Above all, no book is ever the work of a single mind. we remain indebted to the rivers and their diatoms, who continue to remind us that even silence can be full of meaning, and that the smallest voices often carry the greatest truths.

The work presented here is the result of collective effort, shared curiosity, and an enduring passion for understanding aquatic ecosystems through the lens of science and technology.

Our deepest appreciation goes to the academic and research communities devoted to diatom studies and aquatic ecology, whose insights continue to inspire new directions in biomonitoring and environmental conservation. We also acknowledge the developers and contributors of R and DiaThor, whose open-source tools have empowered researchers to transform complex ecological data into meaningful interpretations.

We are grateful to our institutions, mentors, and families for their constant motivation and belief in this endeavor. This book is dedicated to all those who strive to bridge knowledge, innovation, and sustainability for a healthier planet.

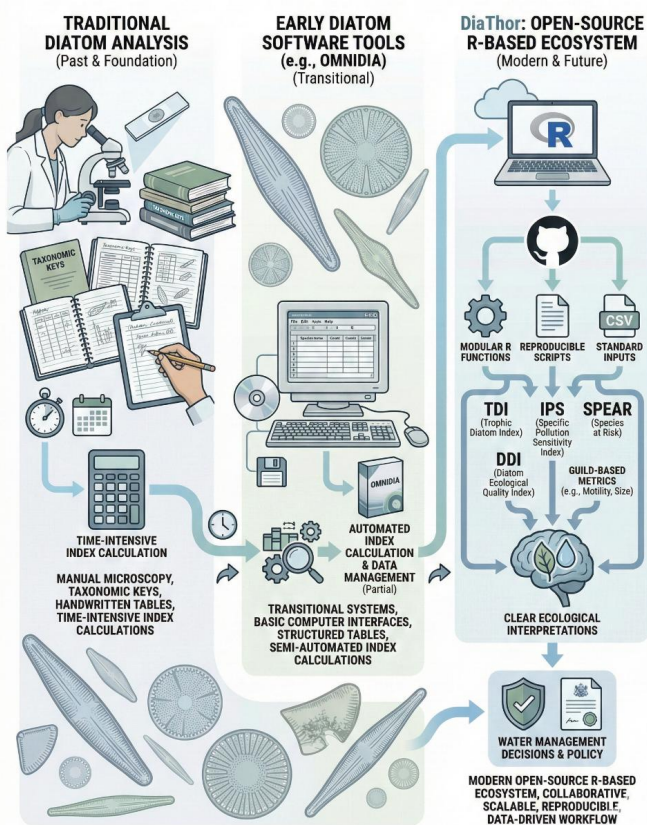
INTRODUCTION

DiaThor is presented as an open-source tool, with its source code available on GitHub, encouraging collaborative development and the incorporation of new functionalities. This open platform allows researchers to suggest and integrate new statistics and indices. Ultimately, DiaThor equips researchers and water managers with a powerful, open-source tool to translate diatom data into meaningful ecological insights. It modernizes diatom monitoring, transforming a labor-intensive process into a streamlined, reproducible workflow. Traditional diatom analysis based on manual identification and index calculation required significant expertise and time.

Early software like, OMNIDIA helped automate parts of this process index calculations and data management. These tools laid the foundation for more advanced platforms like DiaThor, which push Diatom biomonitoring into a more scalable, efficient, and data driven future. Diatoms, a prominent group of microalgae with silica-based cell walls, are known for their sensitivity to ecological conditions (Rimet & Bouchez, 2012). Diatoms respond predictively to environmental stress, making them powerful bioindicators for water quality assessment (Kelly & Whitton, 1995; Wood et al., 2019). Over time, numerous synthesis has emerged like Trophic Diatom Index (TDI), Specific Pollution Sensitivity Index (IPS), and SPEAR-herbicide Index, among others (Kelly & Whitton, 1995; Wood et al., 2019). Additions like the Duero Diatom Index (DDI) and metrics based on ecological guilds and life form strategies allow more nuanced assessments (Álvarez-Blanco et al., 2013; Rimet and Bouchez, 2012). DiaThor integrates many of these within R, offering a flexible and reproducible analysis platform. I'll stress that DiaThor isn't just a "package" you download; it's an ecosystem of functions designed to link taxonomy with ecology. The reason it feels advanced is because it expects you, the user, to already understand diatoms, ecological indices, and R basics. Our task is to bridge those skills into something reproducible and research-ready.

We begin by talking about why diatoms matter. Most of you already know they're not just pretty glass houses under the microscope; they are ecological sentinels. Water quality managers rely on them because they respond quickly to environmental change. That's where DiaThor comes in. Instead of manually calculating ecological indices, a tedious process where mistakes creep in, DiaThor automates the workflow inside RStudio.

EVOLUTION OF DIATOM BIOMONITORING: FROM MANUAL ANALYSIS TO THE **DiaThor** OPEN-SOURCE ECOSYSTEM



Science today demands reproducibility. Your lab mate, or a policymaker, should be able to rerun your analysis and get the same result. That's why DiaThor shines inside RStudio. We'll explore in this book: automating DiaThor pipelines, building parameterized reports, deploying dashboards, so users can interact with data instead of flipping through static tables. This is the difference between a one-off study and a decision-making tool. DiaThor is strict about inputs. If your species names are misspelled or your data table isn't formatted correctly, the functions won't work. That's why you need data discipline, cleaning your species list, standardizing names, keeping raw backups, and always documenting what you changed. Think of it as lab safety, but for your dataset.

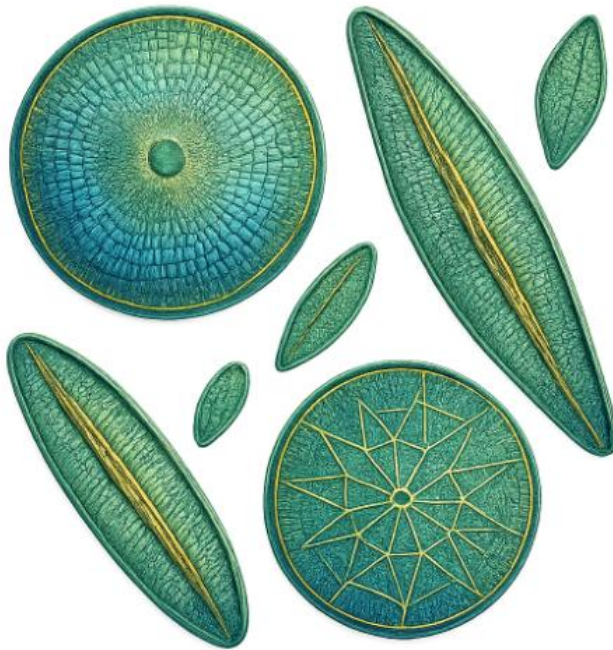
DiaThor can calculate dozens of indices, but should you use all of them? Not necessarily. Each index measures a different aspect of water quality. For example: If your research question is about nutrient pollution, running pesticide indices just clutters your output. Being selective means choosing indices relevant to your study. Being critical means checking whether the results make ecological sense instead of blindly reporting them. Indices give numbers, but traits tell stories. If your results show that motile diatoms dominate, you can say: "This riverbed is unstable, and diatoms are shifting strategy to survive disturbance." That's storytelling. Traits (like Van Dam classifications and guilds) add depth, helping you explain why the index turned out the way it did. It transforms dry tables into narratives about ecosystem health. Numbers in a CSV are hard to interpret. But when you convert guild data into a stacked bar chart or Van Dam traits into a radar plot, patterns jump out instantly. Visualization is how you communicate your findings to non-specialists. Diatoms don't live in isolation; they respond to nutrients, pH, conductivity, oxygen, temperature, metals. If you stop at diatoms alone, your study feels incomplete. By merging DiaThor outputs with water chemistry and hydrological data, you can run ordinations, regressions, or even predictive models. This integration makes your analysis holistic. Instead of saying "Site A has IPS = 14," you can say, "IPS = 14 here because conductivity and nitrate are elevated, and motile guilds are taking over." So, in short, mastering DiaThor isn't about memorizing every function and code. It's about becoming disciplined with data, thoughtful with indices, narrative with traits, creative with visualization, integrative with environment, and reproducible with your workflow.



CHAPTER 1

UNDERSTANDING DIATOMS IN ECOLOGY: THE GLASS-WALLED GUARDIANS

“They are not just indicators – they are integration of time & chemistry holding the blueprint of life in every curve of silica. Big things often come in small packages. And diatoms are living proof of that.”



In the hidden currents of rivers, lakes, and oceans, an invisible world thrives. Minute yet mighty, diatoms are the golden-brown microalgae encased in glass, acting as some of nature's most precise storytellers. These single-celled organisms may be microscopic in size, but their ecological importance is vast, echoing across disciplines from biology to climate science.

Diatoms belong to a group of algae known as Bacillariophyta, characterized by their silica-based cell walls called frustules, which are often exquisitely patterned with intricate symmetry. Each frustule is like a fingerprint, making diatoms not only ecological indicators but also micro-artists of the aquatic world. Frustules are often delicately ornamented. Most are photosynthetic, but some species lack chlorophyll and live heterotrophically among decaying marine algae. Pennate (i.e. bilaterally symmetrical) diatoms occur in both freshwater and marine habitats; centric diatoms (i.e., radially symmetrical) occur predominately as part of the marine plankton. Functionally, they are single cells (the most diverse protists on earth) even though they can appear either in the filaments, chains, or colonies, either in water column (phytoplankton) or attached to any single substratum (benthos). Diatoms possess a remarkable eco-physiological edge when it comes to nutrient acquisition. Among phytoplankton groups, they exhibit significantly high maximum growth rates (Litchman et al., 2006). Diatoms have evolved into one of the most ecologically and biogeochemically important phytoplankton groups. Dominant in nutrient-rich upwelling zones and polar waters, they drive nearly 40% of oceanic primary production (Armbrust, 2009; Tréguer et al., 2018). Their evolution fueled by complex genomic heritage from red and possibly green algal ancestors has given them an extraordinary ability to adapt to changing environments.

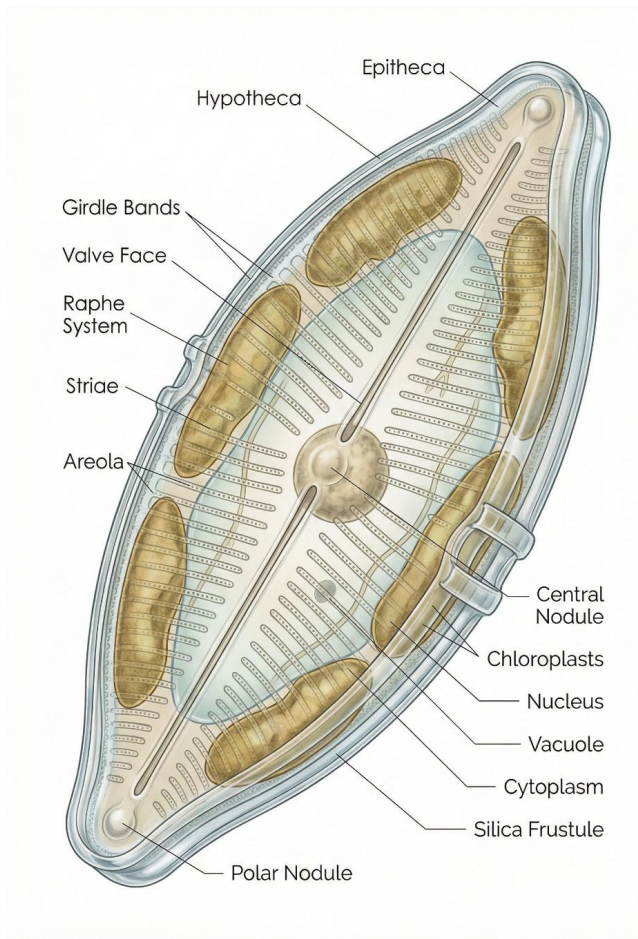


Fig.1.1 Scientific illustration of a diatom showing detailed structural anatomy.

world's rainforests combined. Isn't it amazing that something invisible to the naked eye is quietly keeping us alive? They lock away carbon. Diatoms are climate heroes. When they die, many sink to the bottom of oceans, taking carbon dioxide with them. This helps reduce greenhouse gases in the atmosphere and keeps Earth's climate balanced.

They act as primary agents in the carbon cycle
Diatoms function as primary agents in the biological carbon pump, facilitating the transport and long-term sequestration of carbon to the ocean floor. In aquatic ecosystems, diatoms are at the base of the food chain. Tiny fish, zooplankton, and other creatures munch on them, which then supports bigger fish, birds, and even whales. Scientists love diatoms because they're excellent bioindicators. That means the types of diatoms living in water can reveal whether it's clean, polluted, acidic, or nutrient-rich.

Microscopic examination of aquatic samples reveals the complex structural diversity of diatoms. As photosynthetic microalgae, they construct frustules from amorphous silica, resulting in intricate geometric morphologies. They are microscopic algae, meaning they are single-celled organisms capable of photosynthesis—yes, just like trees, they make their own food using sunlight. Diatoms build their “houses” out of silica (the same stuff in glass). These houses are called frustules, and they come in breathtaking patterns—circles, stars, triangles, rods—almost like nature's version of 3D-printed art. They're everywhere! Diatoms live in oceans, rivers, lakes, even damp soil. Basically, wherever there's water and sunlight, there's a good chance diatoms are busy at work. The next time you walk along a riverbank or see that greenish-brown slimy layer on rocks, that's often diatoms clinging there. They give us oxygen. Here's the surprising part, every fourth breath you take might be thanks to diatoms. They produce about 20–25% of Earth's oxygen, which is as much as all the



DO YOU KNOW?

Recently, scientists have discovered a new freshwater diatom genus named *Indiconema*, in the clean waters of Western Ghats.

(Karthick et al., 2023)

Refer Fig. 1.2 & 1.3

If you sample a river and find pollution-tolerant diatoms, you know the water is struggling. If you find sensitive, delicate species, it's a sign of good health. It's like reading a water-quality diary written by diatoms. Back in the Victorian era, microscope hobbyists would arrange diatom shells into delicate mandala-like patterns, almost like microscopic mosaics.

Today, engineers and nanotechnology researchers study diatom frustules for inspiration. Their designs could help in making better filters, lightweight materials, or even drug delivery systems. The silica shell of a diatom isn't just pretty, it's also practical. It protects them from predators. It allows light to pass through for photosynthesis. It makes them sturdy yet lightweight. Each diatom species has its own shell design, like a fingerprint. That's why taxonomists (the scientists who classify organisms) can identify hundreds of thousands of diatom types. Some are shaped like delicate snowflakes, others like tiny pills, and some look like perfectly drawn geometric patterns. **#Refer Fig. 1.1-1.2**

Diatoms play a crucial role in the biological carbon pump, exporting carbon and silicon to the ocean depths, where it is stored for millennia. Their siliceous cell walls not only contribute to vast marine sediment deposits but also to fossil fuel formation. Genomic studies reveal their unique gene combinations, metabolic flexibility, and ecological dominance, especially in colder, nutrient-rich waters. As climate shifts and ocean chemistry evolves, diatoms remain at the forefront of Earth's carbon cycle and climate regulation, making them key to understanding both past earth systems and future ocean dynamics. (Benoiston et al., 2017).

Nearly all diatoms are microscopic; cells range in size from 2µm to 500µm. The biggest diatoms are about the width of human hair. Scientists use light microscopes (LM) or scanning electron microscopes (SEM) to view diatoms. When diatoms are viewed with a light microscope, the silica walls appear transparent (because we are peering through glass-like structure). When diatoms are viewed with a scanning electron microscope, the cells appear opaque (because we are seeing the surface of the cell with electrons, making the cells appear opaque and textured, offering stunning clarity of their elaborate frustule designs). But their role in ecology extends far beyond productivity.

Diatoms are sensitive to a wide range of environmental variables—nutrients, pH, conductivity, salinity, turbidity, flow rate, and even heavy metals or pesticides. Because different species thrive under different conditions, the composition of a diatom community can be used to infer the ecological status of the waterbody in which they occur.

The ecological value of diatoms lies in their:

Species-specific environmental tolerances (e.g., *Nitzschia palea* thrives in polluted waters; *Achnanthes minutissimum* indicates good quality);

Rapid reproduction rates (allowing them to reflect recent environmental changes);

Widespread distribution (found in freshwater, marine, brackish, and even terrestrial habitats). Because of these traits, diatoms are considered ideal bioindicators—organisms that can be used to assess the ecological condition of an environment. They are particularly useful in biomonitoring programs for rivers, lakes, wetlands, and estuaries, where they help track nutrient loading, pollution events, hydrological changes, and long-term ecological trends. Diatoms are encased in a frustule, a glass-like shell made of amorphous silica (SiO₂·nH₂O). This structure has a high surface area due to its porous and intricate architecture. It bears hydroxyl groups (-OH) and other reactive sites that can bind heavy metal ions through chemisorption.

They act as a physical and chemical sink for dissolved contaminants in the water column. Diatoms secrete sticky mucilage and biofilms, composed primarily of polysaccharides and proteins, which enhance metal ion binding via functional groups (e.g., carboxyl, sulfate). They aid in the adsorption and immobilization of metals within the biofilm matrix. They create microhabitats that trap pollutants and sediments together.

Diatom cell walls act like natural ion exchangers, absorbing metal cations (e.g., Pb^{2+} , Cd^{2+}) by replacing lighter ions like Na^+ , K^+ , or Ca^{2+} . This process is driven by electrochemical gradients and surface charge differences on the frustule. Their preserved frustules can be found in sediments that allows historical analysis as well.

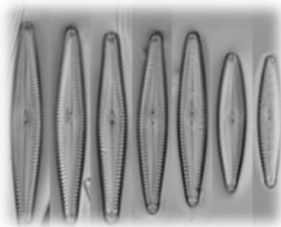


Fig. 1.2 Light microscope images of new genus *Indiconema*



Fig. 1.3 Scanning electron microscope image of new genus *Indiconema*

Some diatoms have the ability to internalize and compartmentalize heavy metals in vacuoles or bind them with organic ligands such as phytochelatins and metallothioneins, thereby reducing their toxicity. Metals are sequestered in vacuoles to isolate them from metabolic pathways. The cellular stress-response proteins detoxify or buffer toxic metals. Because of their fast division rates and wide ecological range, diatom populations can adapt quickly to changing metal concentrations and they can develop resistance or tolerance to sublethal metal stress.

Even polar ice has diatoms! New research from Stanford, published in Proceedings of the National Academy of Sciences, reveals that Arctic diatoms **#Refer Fig. 1.4** microscopic, glass-walled algae found in sea ice, remain motile even at temperatures as low as -15°C (5°F), the coldest known threshold for movement in a eukaryotic cell. Contrary to earlier assumptions that they lie dormant when trapped in ice, the diatoms were observed gliding through microchannels within the ice, propelled by a mucus-like secretion and actin-myosin molecular motors, similar to those driving human muscle movement. The discovery emerged from a 45-day Arctic expedition in the Chukchi Sea aboard the NSF vessel Sikuliaq, where researchers from the Prakash and Arrigo labs employed custom-built microscopes to visualize diatom activity directly inside ice cores. Laboratory experiments recreated icy microenvironments, confirming that Arctic diatoms not only survive but glide faster than their temperate relatives, suggesting evolutionary adaptation to extreme conditions. These traits make diatoms both excellent sentinels of contamination and potential agents for bioremediation. In polluted rivers, lakes, and estuaries, shifts in diatom community structure often reflect increasing metal concentrations. Species such as *Navicula*, *Nitzschia*, and *Gomphonema* are known to dominate in metal-rich environments, making them useful in ecotoxicological studies and pollution index development.

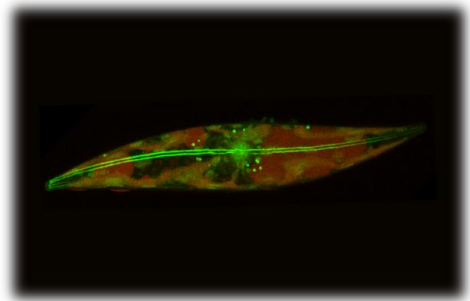


Fig. 1.4 Image of an Arctic diatom, showing the actin filaments that run down its middle and enable its skating motion. | Prakash Lab

You might be wondering: “Okay, diatoms are cool, but what do they mean for me?”

Every sip of clean water may owe something to diatoms, since they help keep aquatic ecosystems balanced. Every breath you take is partly thanks to their oxygen production. Climate stability depends on their role in carbon cycling. And for scientists and students, they are an endless source of discovery and learning. In short, diatoms are not just tiny organisms floating around, they are planet-shapers. The next time you walk by a river or look at the sea, pause for a second. Hidden in that water are millions of diatoms, tiny glass architects, silent oxygen-makers, and storytellers of the aquatic world. They remind us that sometimes the most powerful forces on Earth are not the largest or loudest, but the ones quietly at work in the background.

So, are diatoms interesting?

Absolutely.

They're not just interesting, they're extraordinary.

CHAPTER 2

FROM MICROSCOPES TO MACHINE: THE NEED FOR AUTOMATION

“To automate is not to simplify, it is to standardize complexity with intention”

To effectively manage and analyze the growing volume of biodiversity data generated by these surveys, computational automation has become essential. Understanding the diversity, distribution, and ecological function of diatoms in India has become increasingly feasible, thanks to a combination of government-backed surveys, institutional repositories, and collaborative academic research (e.g., Karthick et al., 2013). Although India lacks a single centralized diatom-based government database, a growing network of regionally focused datasets and national biodiversity platforms provide valuable taxonomic, ecological and distributional data.



The Zoological Survey of India (ZSI) and Botanical Survey of India (BSI) have long documented aquatic biodiversity, including algae and diatoms, as part of their national faunal and floral inventories. In parallel, the India Biodiversity Portal (IBP) and National Biodiversity Authority (NBA) offer searchable, open access data on the occurrence of diatom genera in freshwater and marine ecosystems. These portals serve as semi-governmental platforms, often populated with records sourced from field observations, research publications, and institutional repositories. Several government funded and academic field surveys across different states have yielded rich datasets on diatom community structure:

Rajasthan: Seasonal studies in freshwater lakes have recorded dominant genera such as *Cyclotella*, *Navicula*, and *Nitzschia*, revealing the impact of arid climate on diatom richness.

Jharkhand (Subarnarekha River): A pollution assessment study documented 31 species, including new records for the region, highlighting the effect of heavy metal contamination.

Chhattisgarh: Surveys reported 34 diatom genera across various rivers, with clear seasonal shifts influenced by rainfall and flow variability.

Vindhyan Basin (Uttar Pradesh): One of the largest inventories in north-central India recorded 293 diatom taxa across 49 genera, offering a robust baseline for riverine biomonitoring.

Peninsular India (Urban Wetlands): An IISc-led study covering 43 wetlands identified 181 species, indicating urban nutrient dynamics and their impact on microalgal structure.

Sundarbans Mangrove Sediments: Diatom sampling revealed more than 40 species new to the region, aiding in paleoenvironmental reconstruction.

Gulf of Mannar (Tamil Nadu): Seasonal coastal sampling identified 62 marine diatom species, emphasizing their role in primary productivity and nutrient cycling.

These findings, often supported by state pollution control boards, forest departments, or national missions like the National mission for Clean Ganga (NMCG), demonstrate a decentralized but scientifically rigorous effort in diatom biodiversity monitoring.

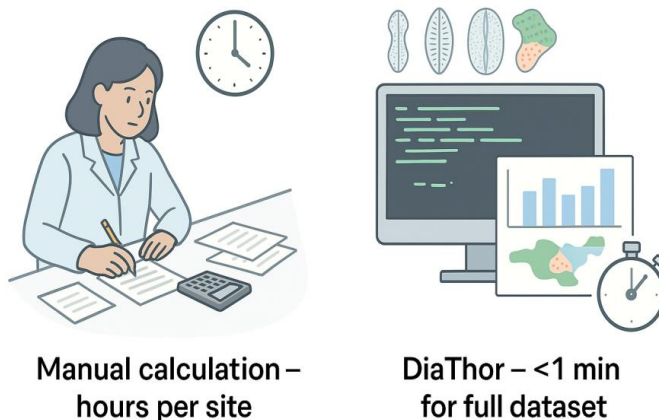
A particularly promising initiative is the Diatom Image Database of India (DIDI), a web-based repository developed through academic collaborations (e.g., IIT-BHU and MJP Rohilkhand University). DIDI offers high-resolution photomicrographs, species-level identification, and ecological metadata for Indian diatoms. However, as these repositories expand, the challenge shifts from data collection to data analysis. To effectively manage and analyse the growing volume of biodiversity data generated by these initiatives, computational automation has become

essential. Consequently, in recent decades, the scientific world has seen a shift from proprietary, black-box software toward open-source ecological tools, a transition driven by the need for transparency, customizability, and community-driven development. These two remove the barriers that once kept ecological modelling in the hands of a few specialists. Instead, they empower students, researchers, and environmental agencies with tools that can be freely modified, shared, and validated across institutions and ecosystems. However, few of these tools match the flexibility, functionality, and user-aligned design of DiaThor. DiaThor is not just another R package, it's a paradigm shifting automated diatom-based biomonitoring.

DiaThor includes a wide suite of globally recognised biotic indices (IPS, TDI, SLA, IDP, SPEAR, DDI, and more), along with trait-based metrics like guild structure, size classes, and chloroplast morphology, all in a single pipeline (Nicolosi Gelis et al., 2021).

With functions like `diat_checkName()` and `diat_loadData()`, it automatically flags mismatches, taxonomic issues and formatting errors and reducing the burden of preprocessing. Built with extensibility in mind, DiaThor allows researchers to add regional indices and expand the internal taxa databases, addressing the geographic bias of earlier European-centric tools. Users can run the master function `diaThorAll()` for everything for selectively execute only what they need, making it ideal for both beginners and advanced users.

Hosted on GitHub, DiaThor's code is transparent, updatable, and backed by a responsive developer community. Contributions are welcome and the tool continues to evolve with user feedback and ecological advancements. By combining scientific robustness, computational efficiency and open-access philosophy, DiaThor marks a new chapter in ecological monitoring. It doesn't just make diatom indexing easier, it makes it shareable, adaptable, and scalable, empowering a generation of environmental scientists to wield better tools for cleaner waters. Diatom-based water quality assessment, traditionally performed through labour-intensive manual methods or expensive commercial tools like OMNIDIA (Lecointe et al., 1993), is now being transformed by open-source platforms that offer both accessibility and reproducibility. These tools remove the barriers that once kept ecological modelling in the hands of a few specialists. While accurate, this approach was time consuming, vulnerable to human error, and



lacked standardization. The early 2000s saw the emergence of semi-automated commercial software like OMNIDIA, which offered built in indices and graphical interfaces for calculating biotic scores. However, these tools often came with steep licencing fees, limited customisation options, and closed-source architecture, restricting scientific transparency, regional adaptability, and collaborative growth.

The increasing volume of diatom data sets generated by ecological monitoring programs, citizen science initiatives, and remote sensing proxies now demands a paradigm shift. The future of bio monitoring calls for tools that are

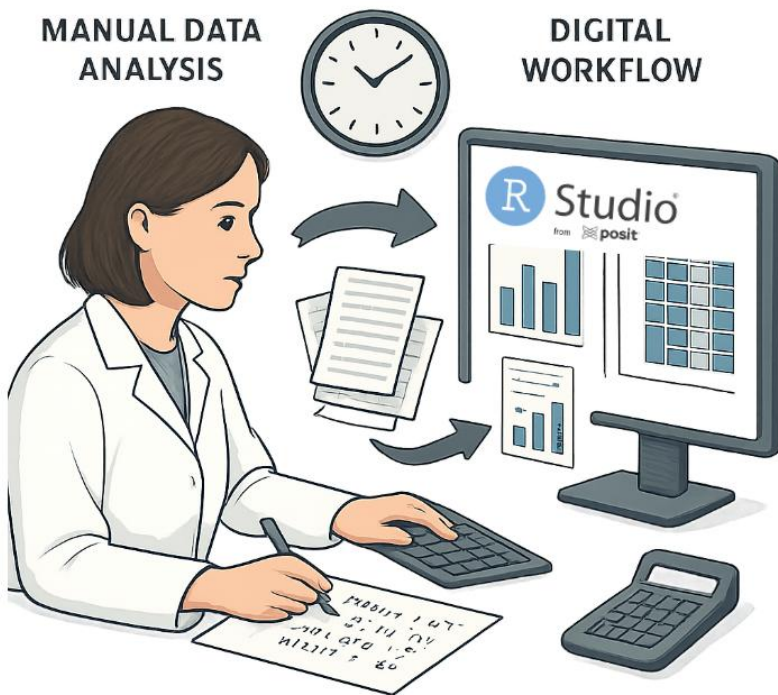
- **Reproducible across labs and studies**
- **Customisable to include local indices and traits**

- **Collaborative for academic, agency, and citizen use**
- **Freely accessible and open to scrutiny**

The need for automation in diatom analysis is driven by more than convenience; it's a scientific necessity. Large scale ecological studies increasingly require high-throughput data processing. Monitoring programs collect data seasonally across dozens of locations, far too much for manual handling. Standardization is essential when results inform policy decisions such as water body classification under frameworks like the European water framework directive or India's clean Ganga mission. Integration with other data types like chemical, hydrological, and genomic requires clean, digitised outputs for downstream modelling. Automation not only accelerates the process but also reduces human bias, improves reproducibility and enables dynamic visualisations for clearer communication with stakeholders. Open-source platforms especially those developed in R, now lead the charge in ecological tool making. These tools promote transparency in scientific workflows, allow the community-driven evolution of new indices, taxonomies, traits etc.,. They enable region-specific adaptation, that is crucial for countries like India where many taxa are underrepresented in Western databases. They foster interdisciplinary collaborations across data science, ecology, and environmental management.

DiaThor, in this context, is not just tool, it is a solution. A bridge between ecological-expertise and data-driven automation, reflecting what modern environmental science aspires to be: fast, open, and collaborative. It's like:

Excel + Statistics + Biomonitoring + Automation, rolled into one.



CHAPTER 3

R STUDIO AND DIATHOR

“In the age of data, the real power lies not in knowing code, but in knowing how to make nature speak through it”



Imagine a damp morning beside a river. You've collected cores, counted diatoms, scribbled GPS coordinates on the back of a notebook. Back home, you open your laptop, that's where R Studio and DiaThor step in and RStudio is where the fieldwork stops being chaos and starts being science. RStudio is not just an editor; it's your stage for reproducible experiments, versioned thinking, and plots that sing.

Think of them as your friendly digital assistants, helping you translate raw data into ecological stories, without needing to be a professional programmer.

First things first: What is R Studio?

You can think of R Studio as the “workbench” where scientists and students organize, clean, and analyze data. At its core, R is a programming language. But R Studio is the user-friendly environment built around R, like a comfortable desk with drawers and tools neatly arranged.

Here, you can type commands, run analyses, and immediately see results in the form of numbers, graphs, or tables.

Now, don't panic at the word “commands.” Using R doesn't always mean wrestling with scary lines of code. With the right packages like DiaThor, you can almost press a few buttons and let the computer do the heavy lifting.

And what about DiaThor?

DiaThor is a special R package designed for diatom analysis. Think of it as a “diatom translator.” You give DiaThor your diatom counts or species list. DiaThor gives you back indices, plots, and interpretations about water quality.

In simpler words, you don't have to know all the complicated ecological formulas. DiaThor already has them built-in, like a calculator that knows every diatom-based index you'll ever need. Why is this useful for biologists? Let's be honest: not every biologist is excited about coding. Many love the fieldwork, the microscope, and the science—but when it comes to software, the enthusiasm drops. That's why DiaThor in R Studio is so powerful. Here's why it's a blessing for non-coders:

- ***It saves time***

Imagine calculating a diatom index manually—species by species, formula by formula. It could take days! DiaThor does it in seconds. You upload your data, run the function, and there it is: the result.

- ***It reduces errors***

Humans make mistakes—especially when juggling large datasets. Computers, on the other hand, follow the formula perfectly every single time. That means more reliable results for your research.

- ***It speaks your language (the ecological one!)***

DiaThor outputs aren't just numbers. It gives you plots, graphs, and tables that clearly show what the diatom community is saying about the environment.

Is the water clean

Is it nutrient-rich?

Is there pollution stress?

DiaThor helps answer these questions visually.

- ***You don't need to be a coder***

Yes, you'll need to learn a few basic commands, but nothing overwhelming. Think of it like learning to drive a car: you don't need to know how the engine works; you just need to know which pedals to press. Once you know the simple steps (like loading data and running DiaThor functions), the package does the "coding magic" for you.

- ***It connects fieldwork to decision-making***

Biologists are often stuck at the stage of raw data. DiaThor takes that data and transforms it into indices that policymakers, environmental managers, or your research guide will actually care about. In short: it makes your science more impactful.

- ***How it feels in practice***

Let's paint a picture.

You're sitting in the lab with your laptop. You've just entered your diatom counts from Yamuna or Betwa River samples. Normally, you'd be stressed thinking, "How do I even start analyzing this?"

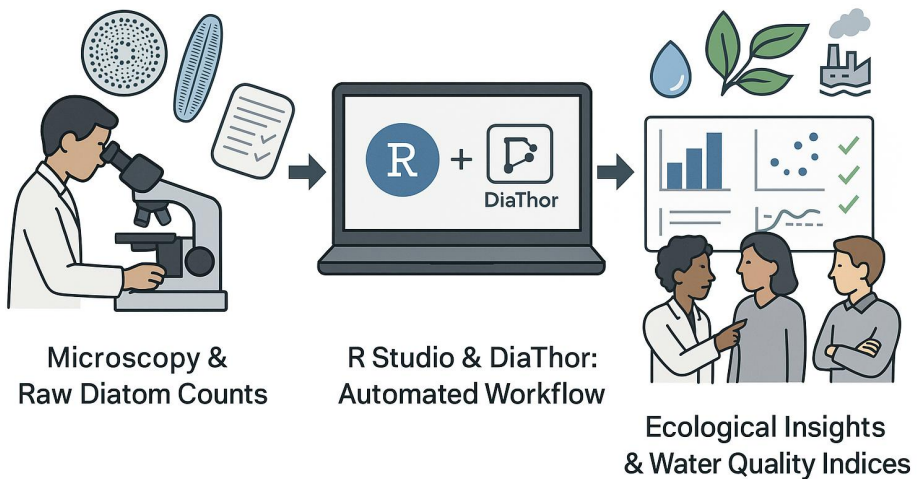
But with R Studio + DiaThor:

1. You, open R Studio (like opening your lab notebook).
2. You load the DiaThor package (like picking the right tool from your kit).
3. You run a simple command (like pressing a button).
4. Within seconds, you have graphs, indices, and reports in front of you.

It's almost like having a smart assistant whispering:

"Look, this site is eutrophic (nutrient-rich). That one shows signs of organic pollution. And here—this stretch of the river is relatively healthy."

Suddenly, your diatoms are not just data points, they're storytellers.



- ***Why this matters in modern biology***

Today, biology is increasingly data-driven. Whether you like it or not, large datasets are here to stay. But the beauty is, you don't need to be a computer scientist to use them. Tools like R Studio and DiaThor are designed to be bridges: Between traditional biology and modern data science. Between your microscope and meaningful environmental insights. For biologists, this means one thing: more time for discovery, less time struggling with calculations. Diatoms already whisper to us about the health of rivers and lakes. But without tools, their voices can be lost in a pile of species counts.

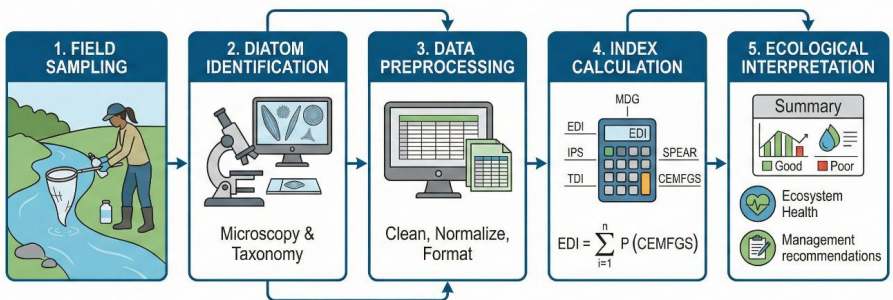
R Studio and DiaThor act like microphones, amplifying those voices so we can hear them clearly. They turn microscopic beauty into ecological truth, and that's a gift for every biologist, coder or not.

While diatom analysis has traditionally required manual computation of indices and careful ecological interpretation, the DiaThor package in R Studio automates much of this process with high accuracy. It incorporates a wide range of biotic indices, allowing researchers to simultaneously evaluate nutrient enrichment, organic pollution, and overall ecological status. Instead of calculating species sensitivities or tolerance values by hand, DiaThor accesses reference datasets, matches species names, and applies standardized formulas, minimizing errors and improving reproducibility across studies.

Another advantage of using R Studio with DiaThor is its ability to integrate multiple datasets and visualize them within a single workflow. Biologists can not only calculate indices but also directly produce graphical outputs, compare water bodies, and link diatom community composition with environmental gradients using ordination methods (e.g., PCA, NMDS). The seamless integration of statistical tools within R Studio means that once the initial data are prepared, advanced ecological analyses, such as diversity measures, cluster analysis, or even machine learning applications, can be layered on top of diatom indices. This makes the workflow highly scalable, from a few river sites in a student project to basin-wide monitoring in professional research.

In essence, R Studio and DiaThor transform raw diatom counts into evidence-based ecological narratives, bridging the gap between classical taxonomy and modern bioinformatics. For biologists, this means less time lost in repetitive calculations and more energy directed toward interpreting patterns, formulating hypotheses, and contributing to conservation decisions. In today's data-driven ecology, such tools do not replace expertise; rather, they amplify it, helping diatoms truly "give voice" to the condition of aquatic ecosystems.

DIA THOR WORKFLOW: Diatom-Based Ecological Assessment



CHAPTER 4

SETTING THE STAGE: R STUDIO AND DIATHOR

“A tool is not knowledge but it becomes powerful when guided by purpose and precision”

To begin installing R on a Windows system, open your preferred web browser and navigate directly to the official CRAN (Comprehensive R Archive Network) website. From the main page, access the section for Windows downloads and choose the latest available version of R. Click the appropriate link to download the installer, then run the executable file and follow the on-screen instructions to complete the installation process. Setting up this foundational statistical environment is the necessary first step before deploying specialized ecological packages like DiaThor (Nicolosi Gelis & Sathicq, 2020) *#Refer Fig. 3.1-3.4*

After successfully installing R, the next step is to set up RStudio, a user-friendly integrated development environment (IDE) for R. Go back to Google and search for “Download RStudio.” Visit the official website at <https://posit.co/download/rstudio-desktop/>, and download the free desktop version available for Windows. Install

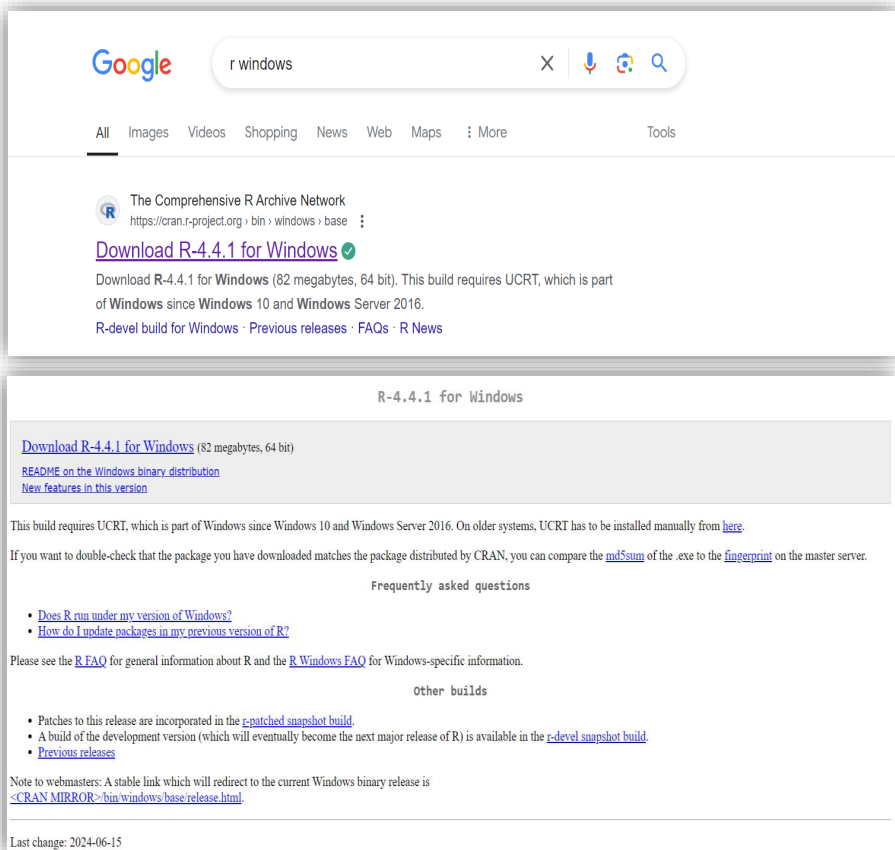


Fig. 3.1 Webpage view demonstrating the process of locating the official download page for RStudio via a Google search.

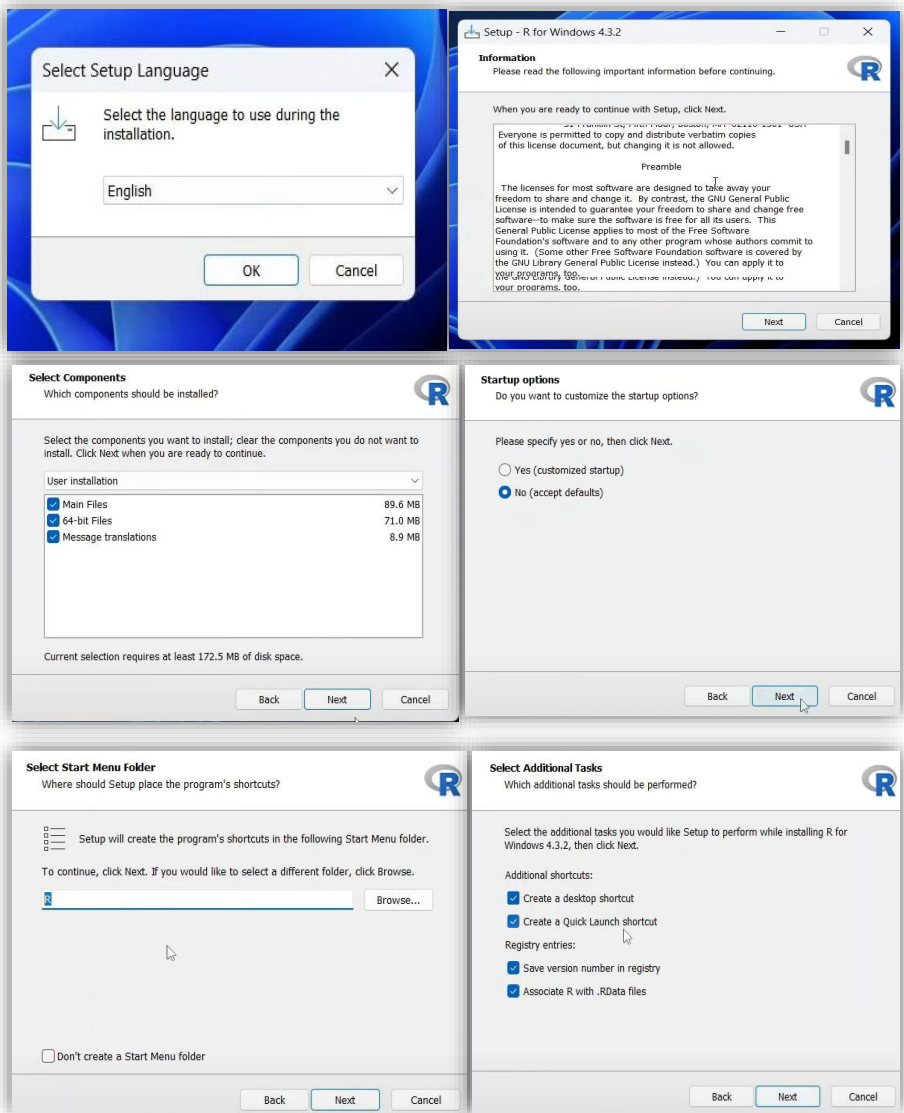


Fig. 3.2 Demonstration of the process of downloading RStudio.

RStudio by running the downloaded file and following the installation prompts. Once installed, open RStudio—you'll see a workspace with several panes, including a Console, Environment, and Script Editor.

To install and run DiaThor, an R package used for computing diatom metrics and ecological indices, focus on the Console pane within RStudio. Type the command `install.packages("diathor")` and press Enter. After the installation is complete, activate the package by typing `library(diathor)` and pressing Enter again. This will load DiaThor and allow you to begin

using its functions for diatom-based water quality assessment. If installed correctly, DiaThor will also appear in the “Packages” tab with a checkmark beside it, indicating it’s ready to use.

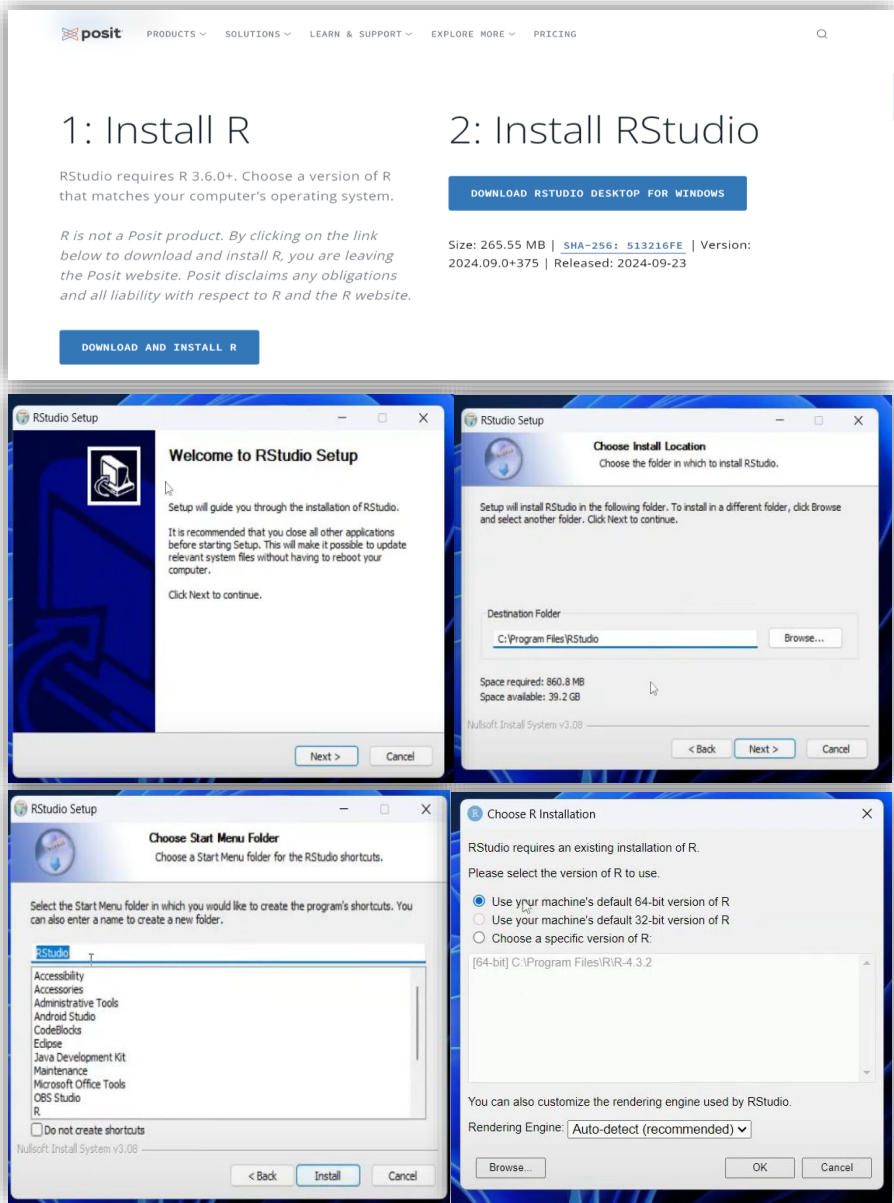


Fig. 3.3 Demonstration of the process of installing RStudio.

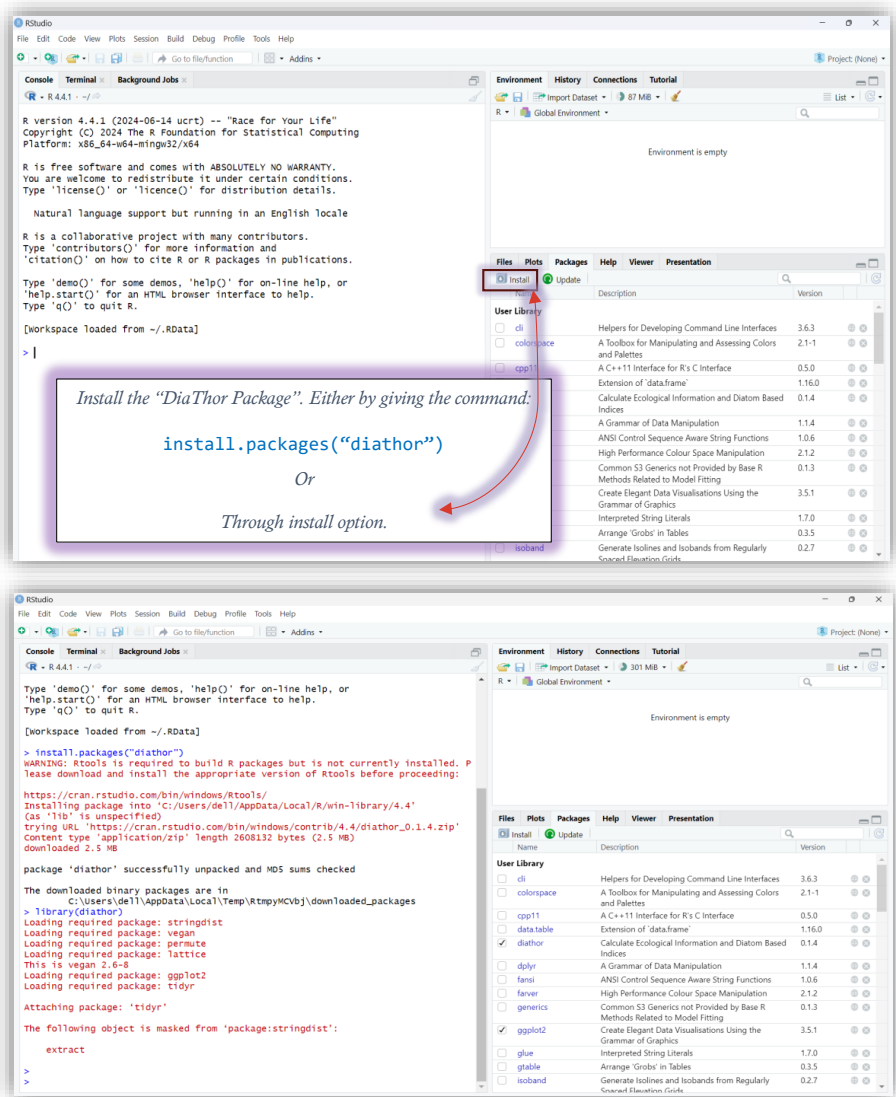


Fig. 3.4 RStudio interface showing the console, script editor, environment, and plots pane; displaying execution of the diathor package for calculating diatom-based ecological indices.

R Studio Interface Breakdown:

Pane	Purpose
Source	Where you write code
Console	Where code executes
Environment	Keeps tracks of your variables
Plots/Files/Help	Shows graphs, directory, or documentation

CHAPTER 5

DATA DICTIONARY FOR DIATHOR INPUT FILES & TROUBLESHOOTING GUIDE FOR DIATHOR

“Data, when structured well, is like a well-packed field backpack, everything in its place, nothing forgotten, ready for the journey ahead.”

Data structuring is a critical prerequisite for reproducible ecological modelling. Transitioning from field sampling to computational analysis requires strict adherence to data formatting protocols. > Encountering error messages regarding missing species names, misaligned columns, or unrecognized objects is a common challenge during the initial stages of data processing. Resolving these issues requires a comprehensive understanding of the data dictionary and the specific input parameters of the DiaThor package. This chapter defines the required structure, permissible values, and formatting rules for diatom input data. DiaThor operates with strict data parsing rules and requires precise formatting to compute indices accurately. The required computational data structure is outlined below.

1. File Format

- **File type:** CSV (.csv)
- **Encoding:** UTF-8 (avoid ANSI or special encodings)
- **Delimiter:** Comma (,)
- **Decimal separator:** Dot (.) not comma

2. Structure

Header Row

- First row must contain **column names**.
- No merged cells or blank headers.

Required Columns

Column Name	Description	Allowed Format	Example
Species	Scientific name of the diatom (must match DiaThor Database spelling)	Genus_species (underscore instead of space), no special characters	<i>Achnanthes_minutissima</i>
Site Columns	Each site/sample has one column	Numeric counts (absolute or relative abundance)	12, 5.7
Optional Metadata	Not required, but may be added if consistent (will be ignored by DiaThor)	Text or numeric	Date, Location, pH

3. Data Types & Rules

- **Species column:**
 - Case-sensitive, must be *exact scientific names*.
 - Use underscore (_) instead of spaces.
 - Avoid extra characters (. , #).
- **Sample columns:**
 - Accepts **integer counts** (absolute abundances, e.g., number of valves) or **relative abundances** (percentages).
 - Must be numeric. Non-numeric entries (e.g., NA, -, present) will break calculations.
- **Missing values:**

- Leave blank ("") or use 0.
- Do not use NA or -.
- **Date (if included):**
 - Use ISO format: YYYY-MM-DD
 - Example: 2025-07-21

4. Example Dataset (**Correct**)

Species	A1	A2	B1	B2
Achnanthes_minutissima	10	12	15	9
Navicula_cryptotenella	5	7	3	8
Gomphonema_parvulum	8	10	6	12
Nitzschia_acicularis	2	4	5	3
Fragilaria_crotonensis	6	8	7	4

5. Incorrect Examples (**Common Mistakes**)

Mistake	Example	Why Wrong
Spaces in species name	Navicula cryptotenella	Should be Navicula_cryptotenella
Special characters	Fragilaria?crotonensis	Not recognized
Non-numeric abundances	present / -	Must be numbers
NA values	NA	Should be blank or 0
Merged headers	A1+A2	DiaThor expects one column = one site



6. Checklist

Before running data into DiaThor, ensure:

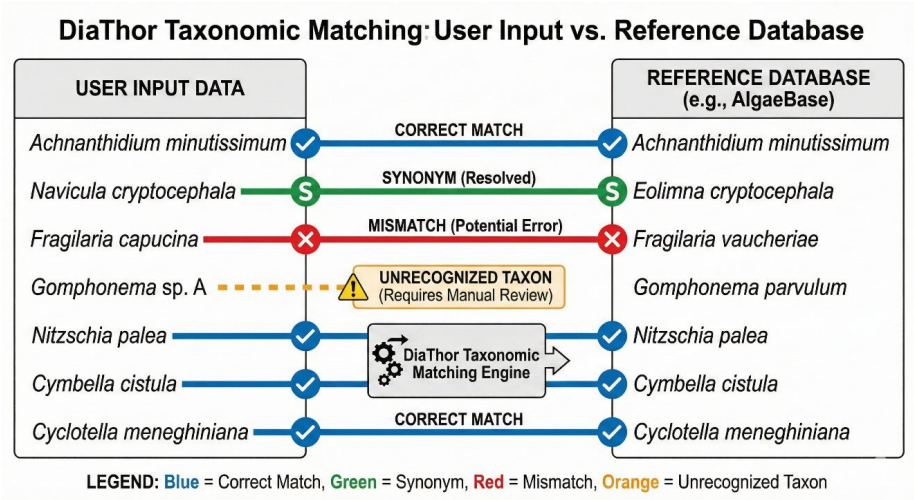
- File is saved as .csv with UTF-8 encoding.
- No merged cells in the header row.
- No blank rows or columns.
- Species names are unique (no duplicates).
- Species names match reference spelling with underscores.
- All abundance values are numeric.
- Decimal separator is a dot (3.5) not a comma (3,5).
- Optional metadata is outside the main matrix (if included).

Adhering to these guidelines ensures successful data loading with the `diat_loadData()` function and minimizes data parsing errors.

Diagnostic Troubleshooting Steps

- 1. Verify CSV Formatting:** The vast majority of execution errors originate from structural or typographical issues within the input file.
- 2. Inspect Data Structure:** Utilize the `str(df)` and `head(df)` commands in the R console to verify the integrity and class types of the loaded data frame.

3. **Isolate the Error:** If a specific index calculation fails, attempt to compute an alternative index. If all index functions return errors, the issue is fundamentally rooted in the input dataset rather than the package functions.
4. **Maintain Software Dependencies:** Ensure all associated packages are current by routinely executing `update.packages()`.



CHAPTER 6

HOW DIATHOR WORKS?

“Application gives knowledge its shape; context gives it meaning, so before we analyze nature, we must first learn to speak the language of data”

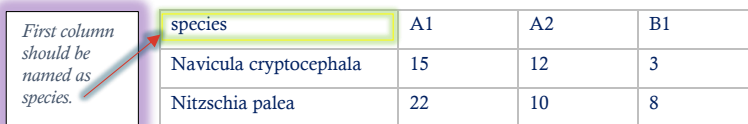
DiaThor is an open-source R package, developed to compute a wide range of **diatom-based biotic indices**, **functional traits**, and **ecological metrics** from taxonomic data. It facilitates standardised, reproducible ecological assessments of water quality using diatom communities as bioindicators. The master function **diaThorAll()** is then run to compute all available indices and ecological metrics in a single process. This function automates the step-by-step calculation of diversity metrics (including richness, Shannon index, and evenness), distribution by size classes, guild composition, and all biotic index values for each sampling site. Internally, it uses sub-functions like **diat_loadData()**, **diat_morpho()**, **diat_size()**, **diat_diversity()**, **diat_guilds()**, **diat_vandam()**, **diat_loadData()**, **diat_ips()**, **diat_tdi()**, **diat_idp()**, **diat_des()**, **diat_epid()**, **diat_idch()**, **diat_ilm()**, **diat_lobo()**, **diat_sla()**, **diat_spear()**, **diat_pbidw()**, **diat_disp()**, **diat_idap()**, **diat_edi()**, **diat_ddi()**, **diat_pdise()**, **diat_cemfgs_rb()**, **diat_checkName()**, **diat_getDiatBarcode()**, **diat_taxaList()**, to generate and compile the full set of outputs. For our dataset, **diaThorAll()** was applied to the diatom abundance data from each site to retrieve the complete range of results. All numerical results were exported to a spreadsheet for deeper inspection and interpretation.

While no advanced statistical modelling was applied in this case, DiaThor outputs are fully compatible with broader environmental datasets

Step 1: Preparing Input Data

DiaThor primarily requires species-level diatom data, typically provided in a UTF-8 CSV format. The first column must be designated for species names, with subsequent columns representing individual sampling sites (e.g., A1, A2) containing absolute counts or relative abundances. Data must be formatted precisely; species names must strictly match the accepted nomenclature within DiaThor's internal database or linked repositories like Diat.Barcode (Nicolosi Gelis et al., 2022). Unrecognized or misspelled taxa will be excluded from subsequent analyses. It is also recommended to use concise sample names to ensure clean, legible labels in the final visual outputs

Example:



species	A1	A2	B1
Navicula cryptocephala	15	12	3
Nitzschia palea	22	10	8

Step 2: Data Loading and Validation

The main data input function is:

```
results <- diat_loadData("your_file.csv")
```

During this step, DiaThor performs an automated validation check against its internal taxonomic databases (Nicolosi Gelis & Sathicq, 2020). The function identifies formatting errors, standardizes taxon names, and flags or removes unrecognized taxa (generating a warning log for user review). If only absolute counts are provided, it automatically calculates relative abundances for downstream index computation.

Step 3: Index and Trait Calculation

Once the taxonomic data is validated, the ecological assessment is executed using the core computational function:

```
result <- diaThorAll(resultsPath="[Enter full output directory path]")
```

This function calculates a comprehensive suite of ecological indicators, including the Trophic Diatom Index (TDI) for nutrient enrichment, the Specific Pollution Sensitivity Index (IPS) for organic pollution, and the SPEAR index for herbicide stress, among others (Nicolosi Gelis et al., 2022). It additionally computes trait-based metrics such as

species	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1
Achnanthes affinis	0	1	1	0	0	0	0	0	0	0	0
Achnanthes bia v bia	20	21	18	22	15	16	21	18	15	19	
Achnanthes clevei	1	0	1	1	0	0	0	1	0	2	
Achnanthes cranulata	2	5	6	2	1	1	5	6	8	4	
Achnanthes exi v exigua	0	2	1	3	1	0	4	0	2	1	
Achnanthes linearis	18	15	12	10	11	12	15	20	12	18	
Ach. Helvetica	0	0	4	2	0	0	3	1	5	1	
Ach marginulata	1	2	0	1	1	0	2	1	4	1	
Achnanthes microcephala	2	1	0	3	0	0	0	4	1	2	
Achnanthes minu v minu	21	18	12	19	15	16	19	23	20	18	
Achnanthes nodosa	0	1	2	1	0	0	1	1	0	3	
Achnanthes pseudoswazi	0	0	0	0	0	1	0	1	0	0	
Achnanthes rivulare	1	0	2	0	1	1	0	0	3	1	
Achnanthes subhudsonis	1	0	0	0	0	1	0	0	0	1	
Adlafia miniscula (Grunow) Lange-Bertalot	3	2	0	2	1	0	3	1	0	4	
Adlafia muscora (Kociolek & Reviere) Lange-Bertalot	0	0	2	1	0	0	2	3	0	1	
Amphora copulata	2	1	0	1	0	2	1	1	0	2	
Amphora Inariensis	0	1	0	1	0	0	1	0	1	0	
Amphora ovalis	3	2	1	1	1	3	2	1	1	1	
Amphora pediculus	4	2	1	5	6	2	1	1	5	6	
Amphora veneta	0	2	2	1	0	3	1	2	0	0	

This is a sample file that must contain the first column named "species" in lowercase letter, and sampling sites named as A1, A2, etc. as highlighted with yellow color in the picture above.

Then, it has the diatom relative abundance data count and the species name written correctly without any spelling mistakes and without repeating any species name

Fig. 5.1 Structure of the input CSV file for DiaThor analysis, showing species names and corresponding abundance values.

ecological guilds, size classes, and saprobity classifications. Each index is computed automatically using species relative abundances cross-referenced with the sensitivity scores and environmental preference values built into the DiaThor ecosystem. This triggers the calculation of:

- **Diatom indices** such as:
 - ✓ Trophic Diatom Index (TDI)
 - ✓ Specific Pollution Sensitivity Index (IPS)
 - ✓ SPEAR herbicide Index
 - ✓ Descy Index (IDG)
 - ✓ Sládeček Index (SLA)
 - ✓ Intercalibration Common Metrics (ICMi)
 - ✓ Duero Diatom Index (DDI)
- **Ecological guilds** (low-profile, high-profile, motile, planktonic)
- **Size classes**
- **Life forms**
- **Chloroplast number and shape**
- **Biovolume estimates**
- **Saprobity classes** (organic pollution tolerance)

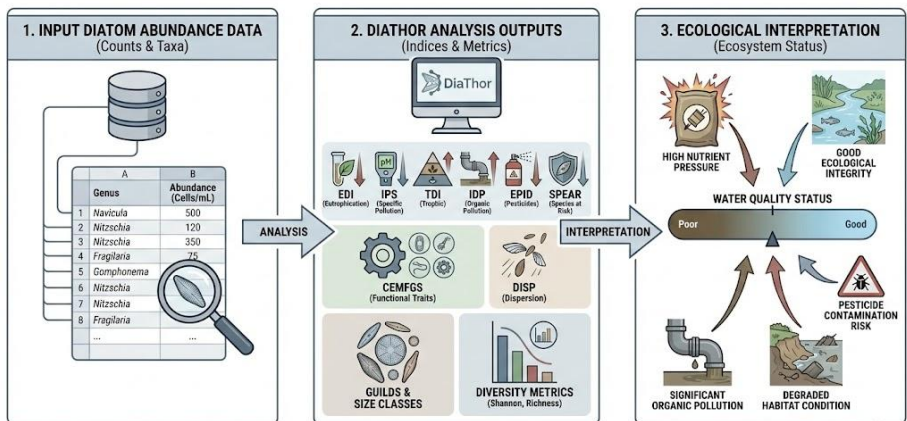
Each index is computed based on species presence/abundance, sensitivity scores, and environmental preference values stored in DiaThor's internal datasets.

Step 4: Visualization and Outputs

If the `plotAll = TRUE` argument is passed, DiaThor automatically generates graphical summaries of the ecosystem's health, including bar plots of index values across samples, guild composition charts, and biovolume distribution graphs. All computational results are exported as .csv tables containing the numerical index values, alongside PDF reports with the summary plots. However, ecological interpretation remains the responsibility of the researcher; the generated numerical outputs must be contextualized with local environmental data to determine the specific stressors affecting the water body.

Step 5: Advanced Modular Use

While `diaThorAll()` provides a generalized overview, researchers can also extract or calculate specific components using modular functions like `diat_ips()`, `diat_tdi()`, or `diat_guilds()`. This modularity allows for targeted, hypothesis-driven assessments and seamless integration with other environmental datasets for advanced statistical modelling (e.g., regressions, ordination) using supplementary R packages like `ggplot2` and `vegan`.



CHAPTER 7

VOICE OF THE WATER: THROUGH THE LENS OF DIATOM INDICES

“Each Index is a lens, and what we see depends on how clearly we look through it”



Biomonitoring using diatoms relies on translating biological data into interpretable metrics. These metrics are known as Ecological Indices, that provide a quantitative assessment of water quality based on the composition and abundance of diatom species.

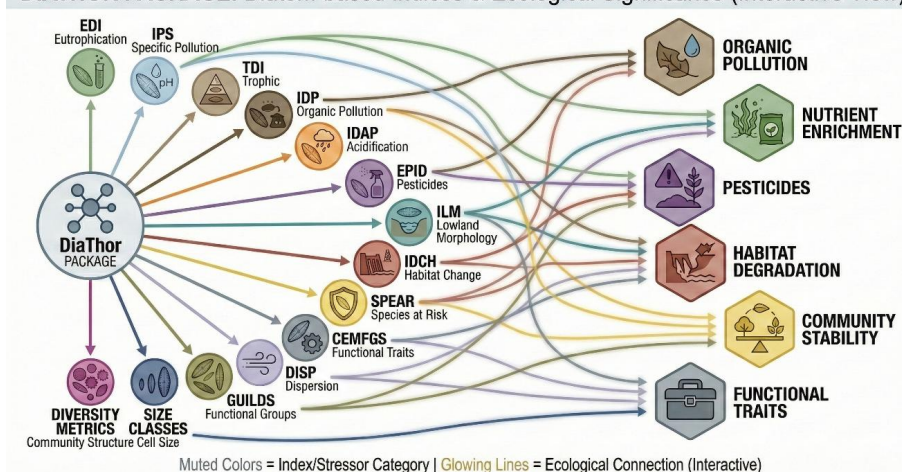
DiaThor allows users to compute multiple indices automatically, offering a rich framework to evaluate ecological conditions across time and space. However, to interpret the outputs meaningfully, it is essential to understand what each index measures, how it is calculated, and what ecological insight it provides.

Indices are standardized formulas that transform biological community data (for e.g., species composition, abundance, sensitivity scores) into a single value reflecting ecological health. These are often based on:

- **Species tolerance or sensitivity**
- **Pollution gradients**
- **Nutrient levels**
- **Oxygen demand**
- **Organic enrichment**

They provide a numerical simplification of complex ecological patterns, useful for comparison, regulation, and environmental reporting. However, their power lies in the context, not just the number. When used carefully, they reveal how microscopic life forms like diatoms mirror macroscopic environmental change. In the upcoming chapters, we'll bring these numbers to life through visualizations that allow pattern recognition at a glance. Understanding each index, what it measures, the data it requires, how it is calculated, is essential to interpreting the biological meaning behind the numbers it outputs.

DIATHOR PACKAGE: Diatom-based Indices & Ecological Significance (Interactive View)



1. **IPS (Indice de Polluo-Sensibilité Spécifique)**- An index developed in France (Coste, 1982; Cemagref, 1982) that indicates general water quality based on pollution-sensitive vs. tolerant diatoms. IPS ranges roughly from ~1 (heavily polluted) to ~20 (very clean waters), with higher values indicating better water quality (more sensitive, clean-water species). The IPS (Specific Pollution Sensitivity Index) is one of the most widely accepted metrics

for assessing water quality using diatoms. It is particularly sensitive to organic pollution. To calculate IPS, the software requires species names and their relative abundances. Each species has an associated sensitivity value (S) and an indicator value (V), both derived from ecological literature. The formula used is:

$$IPS = \sum(A \times S \times V) / \sum(A \times V)$$

where A is the relative abundance of the species. The index ranges from 1 (very poor water quality) to 20 (excellent water quality). The output also includes the percentage of taxa used in the calculation and the number of sensitive taxa contributing to the score.

2. **TDI – Trophic Diatom Index**- Developed in the UK to reflect nutrient enrichment, especially phosphorus levels (Kelly & Whitton, 1995). Like the IPS, it requires species names and relative abundances. Each species is given a sensitivity score related to its tolerance to eutrophication. The index ranges from 0 to 100, with higher values indicating increased nutrient levels. The formula is:

$$TDI = \sum(\text{Abundance} \times \text{Sensitivity}) / \sum(\text{Abundance})$$

The TDI is especially useful in detecting the effects of agricultural runoff and urban wastewater on aquatic systems.

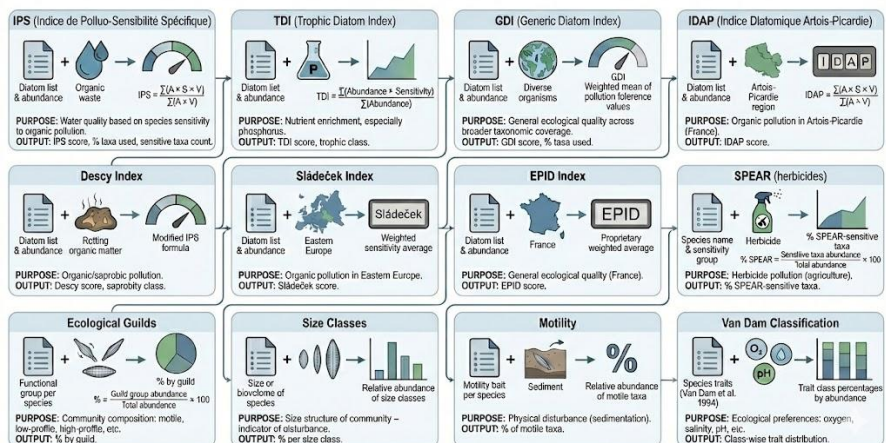
3. **GDI – Generic Diatom Index**- It was designed to serve as a more inclusive alternative to IPS Coste & Ayphassorho (1991). It covers a broader range of diatom taxa and is especially helpful when dealing with incomplete taxonomic datasets. It also uses species names and abundances, and calculates a weighted mean of pollution tolerance values. The GDI is a generalized water quality indicator and is well-suited to large-scale surveys where traditional indices like IPS might not provide full coverage.
4. **IDAP – Indice Diatomique Artois-Picardie**- The IDAP is a regional French index specifically used in the Artois-Picardie watershed (Prygiel et al., 1996). It measures the same parameters as the IPS but uses slightly different sensitivity and indicator values tailored for local species. The calculation formula is similar:

$$IDAP = \sum(A \times S \times V) / \sum(A \times V)$$

It provides water quality categorization based on regional ecological knowledge.

5. **Descy Index**- Developed by Jean-Pierre Descy, this index also targets organic pollution and is similar in concept to IPS. It uses species abundances and predefined saprobic values for each species. While the formula resembles IPS, it is modified to suit different regional and ecological contexts. The index helps assign water bodies to saprobic classes and is useful for comparative studies in Central and Western Europe. (Descy, 1979)
6. **Sládeček Index**- This index is used primarily in Eastern Europe and is built upon species' saprobity scores. It requires species names and abundances, and calculates a weighted average based on how tolerant each species is to organic enrichment. It provides a complementary perspective to IPS and is especially relevant when assessing older datasets or working in legacy monitoring systems. (Sládeček, 1986)
7. **EPID – Indice EPIDémiologique**- The EPID index is a lesser-known French index used for general ecological assessment. It works similarly to the IPS but is based on a different set of sensitivity values and

DIATHOR INDICES AND OUTPUTS: ECOLOGICAL ASSESSMENT FRAMEWORK



classification rules. It again relies on species abundances, and the result is a simplified water quality score derived from proprietary formulas. This index is best used for local or complementary monitoring in French regions. (Dell'Uomo, 1996).

8. **SPEAR(herbicides)**- Unlike the previous indices, the SPEAR (Species At Risk) index is focused on herbicide pollution (Wood et al., 2019). It classifies diatoms based on their sensitivity or resistance to pesticides, particularly herbicides. The input required includes species names and their sensitivity group (sensitive or tolerant) (Wood et al., 2019). The formula is:

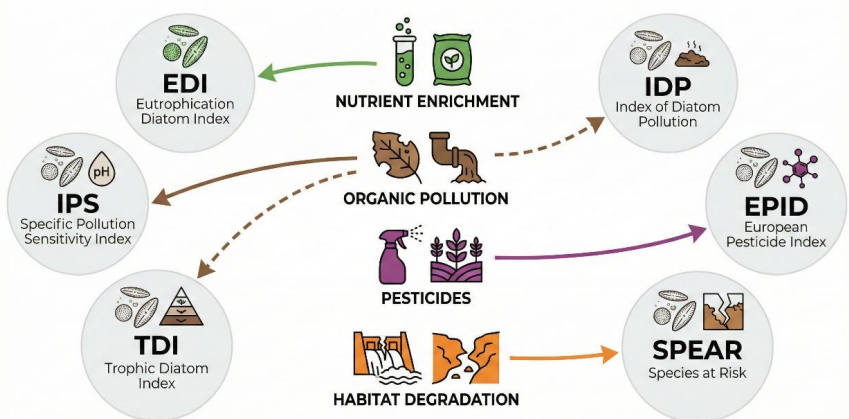
$$\% \text{ SPEAR} = (\text{Abundance of sensitive species} / \text{Total abundance}) \times 100$$

This index is useful in agricultural landscapes where pesticide runoff is a concern.

9. **Ecological Guilds**- This output provides insight into the functional composition of the diatom community. Each species is categorized into ecological guilds like low-profile, high-profile, motile, and planktonic, based on their growth form and behavior. The required input is the guild assignment per species, which DiaThor pulls from internal databases. It calculates the percentage abundance of each guild, giving insight into habitat stability and disturbance. For example, a dominance of motile taxa often reflects unstable or disturbed environments (Passy, 2007; Rimet & Bouchez, 2012).
10. **Size Classes**- Size class analysis involves categorizing species into predefined biovolume classes (e.g., very small, small, medium, large). It reflects ecological strategies such as competition, light penetration, and grazing resistance. The data required include species names with known size class information. DiaThor then computes the percentage of species in each size category. (Litchman et al., 2009) This functional trait helps understand shifts in community structure due to environmental stress or successional changes.
11. **Motility Classification**- Motility is another functional trait, with species being categorized as motile or non-motile. Motile taxa can move across surfaces and are often found in environments with high sedimentation or frequent disturbance. DiaThor calculates the relative abundance of motile species, providing insight into the physical dynamics of the aquatic habitat.
12. **Van Dam Classification**- This classification is based on the environmental preference system developed by Van Dam et al. (1994). It assigns diatom species into classes based on several environmental variables such as:
 - Oxygen requirements (aerophilic to anaerobic)
 - pH preference (acidic to alkaline)
 - Salinity tolerance (fresh to brackish)
 - Saprobity (pollution sensitivity)
 - Nitrogen uptake metabolism

DiaThor outputs the percentage of species falling under each category within each parameter, allowing multi-dimensional ecological interpretation. It requires only species names, as all trait data is sourced from DiaThor's built-in database.

DIATOM-BASED INDICES & ENVIRONMENTAL STRESSORS



THE FUNDAMENTAL MECHANICS OF DIATOM INDICES:

The calculation of ecological indices relies on two primary variables:

- **Species Identity and Sensitivity:** Individual diatom taxa exhibit varying degrees of tolerance to environmental stressors. Sensitive taxa thrive exclusively in unpolluted waters, whereas tolerant taxa can proliferate in degraded conditions.
- **Relative Abundance:** The proportional representation of these taxa within the community dictates the final index score

Stand beside any river long enough, and you'll notice something: water never speaks in words, but it always tells stories. The shimmer on its surface, the earthy smell after rain, the slow bend around a stone, all are fragments of a hidden language. But when we, as scientists, want to truly decode the water's voice, we often turn to its tiniest residents: the diatoms.

Yes, those microscopic algae with glassy shells. They're the stenographers of aquatic life, writing down every chemical shift, every nutrient surge, every brush with pollution, except they don't use ink. They record it in their presence, abundance, and community structure.

Why Diatoms?

Imagine water as a grand orchestra. You might think fish are the violins, macrophytes the percussion, and crustaceans the brass. But diatoms? They're like the tuning forks, exquisitely sensitive, shifting instantly when the composition of water changes.

Give them extra nutrients? They flourish, sometimes too much. Lower the oxygen? Sensitive species disappear, hardy ones move in. Pump in heavy metals or detergents?

Their shells, their community, their very survival begins to shift.

In short: diatoms don't lie. If water has a secret, diatoms spill the tea.

The Birth of Indices

Now, here's where the story gets numerical. Scientists realized long ago that simply looking at species wasn't enough. We needed a way to translate the diatom chorus into a quantifiable score. And so were born the diatom indices, mathematical recipes that take your diatom list and spit out a number reflecting the water's health.

It's like taking a messy diary and turning it into a headline:

IPS: "Pollution sensitivity: 15/20."

TDI: "Nutrient load: trending high."

Saprobity Index: "Organic matter level: yikes, too much!"

Each index is basically a different lens. Put one lens on, you see nutrients. Switch it, and you see oxygen. Another one, and you see heavy metals. Together, these lenses give you a complete portrait of the river's wellbeing.

How the Indices Work (Without the Jargon)

Let's keep this light. An index is basically:

Species identity matters. Some diatoms are sensitive (they only thrive in clean water), while others are tolerant (they shrug at pollution like, "meh").

Abundance matters. A single rare species won't sway the score. But if tolerant species dominate, that's a red flag.

Math magic happens. Weighted averages, reference scores, sensitivity values—all churned together into one tidy number.

So, one number (say, an IPS score of 18) is shorthand for: "Hey, this water is in good shape, lots of clean-loving species here."

Picture this:

A mountain stream, cold and bubbling. Diatom indices here will sing like a clean flute: high IPS, low TDI, low saprobity. Translation? "Pure, oxygen-rich, keep me this way."

A city canal, lined with factories. Diatoms here change their tune, motile, nutrient-tolerant species dominate. Indices read: low IPS, high TDI. Translation? “Too much fertilizer, too much organic waste. I’m choking.”

A river recovering after cleanup. The indices slowly climb, diatom communities diversify, sensitive species sneak back in. Translation? “Thanks, I can breathe again.” It’s not just data; it’s water writing poetry in numbers.

DiaThor: The Interpreter

And here’s where our friendly R package, DiaThor, enters the chat. Instead of us scribbling formulas in lab notebooks, DiaThor acts as the interpreter between diatoms and humans. You feed it your species list; it feeds you a buffet of indices and graphs.

It’s like hiring a professional translator for a river’s diary.

But remember: even with DiaThor, interpretation still rests with you. The software gives you the numbers, but it’s your ecological wisdom that makes sense of the song.

Closing: Listening Better

Diatoms don’t just sit in petri dishes or microscope slides. They are the living echo of water’s voice. Every shell, every cell, every shift in abundance is a syllable in the ongoing conversation between nature and human impact.

When we calculate indices, we’re not reducing rivers to cold math. We’re amplifying their whispers into a language policymakers, engineers, and communities can hear.

So next time you dip into a stream, think of it this way: the water is already speaking. Diatom indices are just the subtitles.

Do you know what is Saprobic value or Saprobity?

"How dirty (organically) is this water, and who's thriving in it because of that filth?"

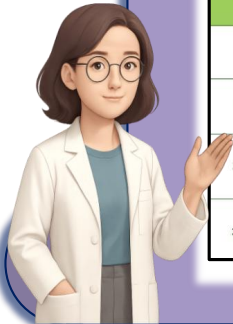
Saprobic value, also known as saprobity, is a biological indicator used in aquatic ecology to assess the degree of organic pollution in water bodies such as rivers, lakes, and streams. This concept is based on the ecological principle that different species of aquatic organisms, particularly diatoms, protozoa, macroinvertebrates, and bacteria, have varying tolerances to organic matter—especially decomposing material like sewage or plant detritus. Some species are highly sensitive and only thrive in clean, oxygen-rich water, while others are more tolerant and flourish in nutrient-rich, polluted conditions. By identifying which organisms are present in a water sample and how abundant they are, ecologists can infer the level of organic pollution that water has been subjected to over time.

Saprobity is expressed as a numerical index, typically ranging from 1.0 (very clean) to 4.0 (heavily polluted). Lower values indicate oligosaprobic conditions, where only pollution-sensitive species survive, suggesting excellent water quality. Mid-range values suggest mesosaprobic conditions, often divided into beta (moderately polluted) and alpha (highly polluted) zones. Values near 4.0 indicate polysaprobic conditions, where only pollution-tolerant organisms can persist, reflecting severe organic degradation and low oxygen availability. These values are not arbitrarily assigned; rather, each species is given a specific saprobic value based on long-term ecological studies, and a site's overall saprobity is calculated using a weighted average of the species present and their abundances.

The saprobic system was pioneered in the early 20th century by Kolkwitz and Marsson and later refined by Pantle and Buck, and by Sládeček. It became one of the most widely used biological methods for evaluating water quality in Europe and globally. One of the key strengths of saprobity indices is that they reflect long-term environmental conditions rather than just the current snapshot, unlike chemical tests that can miss episodic pollution events. Moreover, saprobic assessments are relatively cost-effective and don't require complex lab instrumentation—just skilled identification of indicator organisms.

However, saprobity is best suited to assessing organic pollution, such as untreated sewage or decaying plant matter, and may not accurately capture impacts from toxic pollutants, heavy metals, or thermal discharges, unless the species' response to those factors is well-documented. In modern aquatic biomonitoring, saprobic values are often integrated with other biological indices to give a fuller picture of ecological health. Tools like the DiaThor R package automate the calculation of saprobity and other diatom-based ecological traits, making it easier than ever for researchers and students alike to apply this classic, yet still powerful, concept in environmental science..

Saprobic Value	Water Quality	Organic Pollution Level	Common Taxa
1.0–1.5	Oligosaprobic	Clean, low pollution	Sensitive diatoms like <i>Achnanthes minutissimum</i>
1.5–2.5	β-Mesosaprobic	Moderate pollution	Medium-tolerant diatoms
2.5–3.5	α-Mesosaprobic	High pollution	Species like <i>Nitzschia palea</i>
3.5–4.0	Polysaprobic	Gross, highly polluted	Sludge-loving species



CHAPTER 8

WHAT IS A TIDY FORMAT?

“Messy data is a story half-told; tidy data is a story ready to be read.”

In simple terms, a **tidy data frame** is a well-organized table where each column is a variable, each row is an observation, and each type of observation unit forms a table. This concept was formalized by Hadley Wickham, a key contributor to the R ecosystem.

Why This Structure Matters:

Tidy data frames are foundational to reproducible and flexible scientific analysis. They allow you to perform tasks such as:

- **Quickly generating figures with minimal code,**
- **Filtering for samples with low-quality scores,**
- **Comparing results across indices or trait groups,**
- **Exporting results to CSV, Excel, or interactive dashboards.**

More importantly, tidy data aligns with modern data science practices in R. Nearly all tidyverse functions (from packages like `dplyr`, `ggplot2`, `tidyr`, etc.) are optimized for tidy data. When working with DiaThor, understanding and maintaining this structure ensures that your outputs are immediately ready for downstream analysis and visualization

Characteristics of a Tidy Data Frame:

Rule	Explanation
Each column is a variable	e.g., sample name, IPS score, % taxa used, etc.
Each row is an observation	e.g., one row per sample
Each cell contains a single value	No multiple values crammed into one cell
Each type of data is in one table	You don't mix unrelated data in one table — each table is about one unit type

CHARACTERISTICS OF A TIDY INDEX DATA COLUMN:

Index	Tidy Data Columns
IPS Index	SampleID, IPS, TaxaUsed_Percent, SensitiveTaxa_Count
TDI Index	SampleID, TDI, Trophic_Class
GDI Index	SampleID, GDI, TaxaCoverage_Percent
Descy Index	SampleID, Descy, Saprobity_Class
Sládeček Index	SampleID, Sladeczek, Class
IDAP Index	SampleID, IDAP
EPID Index	SampleID, EPID
SPEAR Index	SampleID, SPEAR_Percent
Van Dam Classification	SampleID, TraitType, TraitClass, Relative_Abundance
Ecological Guilds	SampleID, Guild, Relative_Abundance
Size Classes	SampleID, Size_Class, Relative_Abundance
Motility	SampleID, Motility_Type, Relative_Abundance

Imagine you're back from the field, exhausted but excited, with samples, notes, and numbers spilling out of your backpack. You've got diatom counts on crumpled papers, water chemistry in one Excel sheet, and site codes written in a notebook. That's what your raw data looks like: messy, patchy.

And that's fine, science is born messy. But when it comes time to analyse? Computers are divas. They don't like messy. They want order, structure, discipline. That's where the idea of tidy format steps in.

What Exactly is "Tidy Format"?

Tidy format is just a fancy way of saying: each piece of data has its place, and everything is lined up neatly.

Specifically:

Each variable gets its own column.

Example: "Species," "Abundance," "Site," "Date."

Each observation gets its own row.

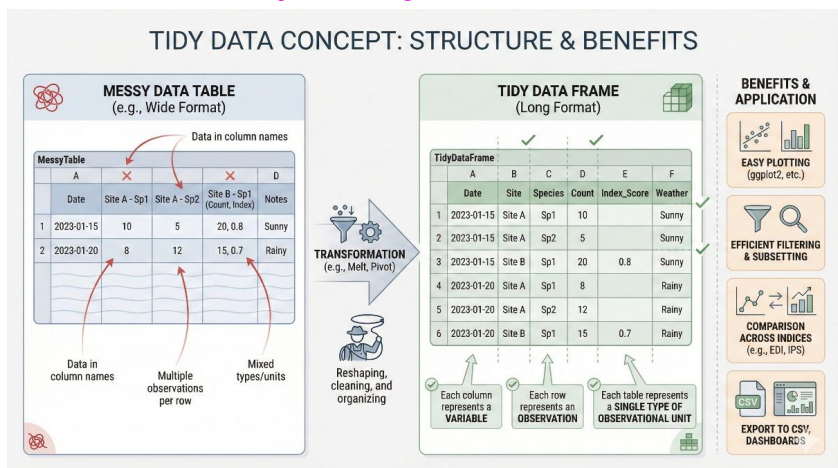
Example: One line = one species counts at one site on one date.

Each type of data gets its own table.

Example: Don't mix diatom data with pH readings in the same table, give them their own tables, but keep the structure tidy.

Why Biologists Should Care?

You're not just cleaning tables for the fun of it.



Tidy data means:

Easier analysis; Most R functions (like those in the tidyverse) are designed to work with tidy data.

Reusable code; You don't have to rewrite your scripts for every dataset.

Visualizations that actually work. ggplot2 loves tidy data—it's like peanut butter meeting jelly.

Collaboration; A tidy dataset is like a neatly written lab notebook. Anyone can pick it up and understand it.

Tidy data is like arranging your spice rack. If cumin is in one jar, coriander in another, and each jar has a clear label, you can cook a masterpiece without wasting time. But if all your spices are dumped into one bag? Good luck finding the cinnamon.. So, the next time you wrangle your tables into tidy format, just remember, you're not just cleaning data. You're unlocking its voice.

CHAPTER 9

ESSENTIAL R PACKAGES FOR ECOLOGICAL VISUALIZATION

“Visualization is the bridge between data and discovery”

Before we dive into the world of ecological visualization in RStudio, let’s pause and set the scene; it’s important to recognize that R is not a monolithic tool.. R is not just a single tool or monolithic software; it’s more like a modular ecosystem, a Swiss Army knife of statistical and graphical possibilities. Its true magic comes alive when we load the right packages, the specialized toolboxes that let us perform precise tasks: plotting, statistical transformations, mapping, or exploring complex ecological datasets. Instead, it’s a **modular ecosystem**. Each package is a specialized toolbox, offering functions that let you create specific types of plots, conduct statistical transformations, or fine-tune visual aesthetics.

When it comes to visualizing DiaThor outputs or any ecological dataset teeming with species, traits, and environmental indices, having the right packages is essential. These packages span basic plotting, advanced graphics, spatial mapping, ordination, and even interactive dashboard creation. Let’s take a guided tour through these packages, why they matter, and how they transform raw data into clear ecological stories.

To visualize DiaThor output—or any complex ecological dataset—you’ll need to install and load a suite of packages that span basic plotting, advanced graphics, spatial visualization, and multivariate ordination. Below is a breakdown of the **essential packages**, organized by their purpose, along with detailed descriptions of what they do and why you need them. *#Refer Fig. 8.1 & 8.2*

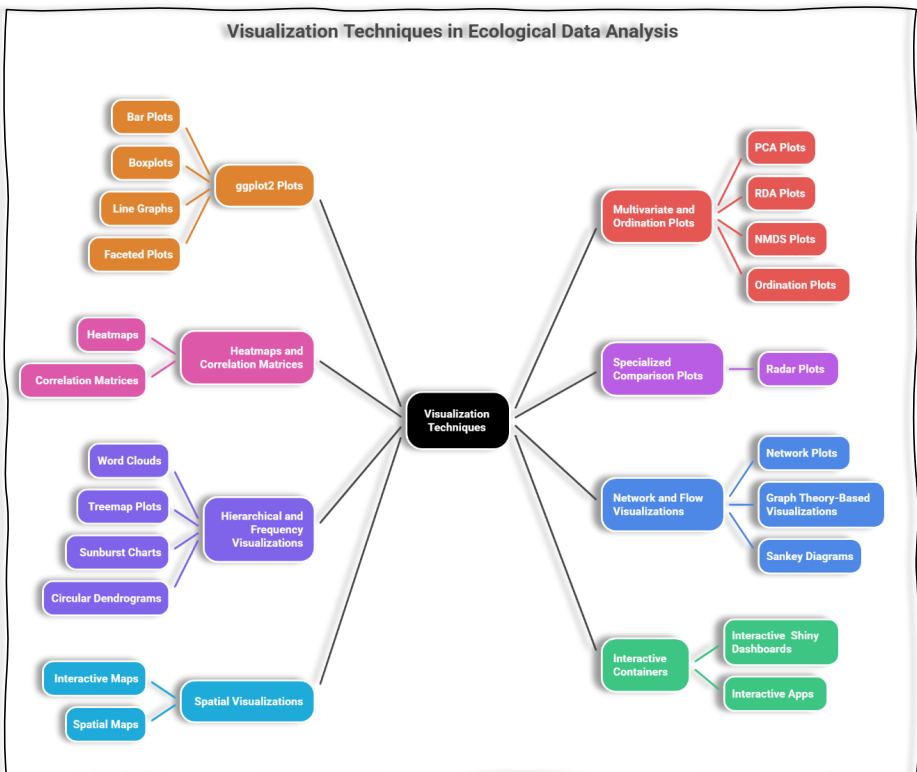
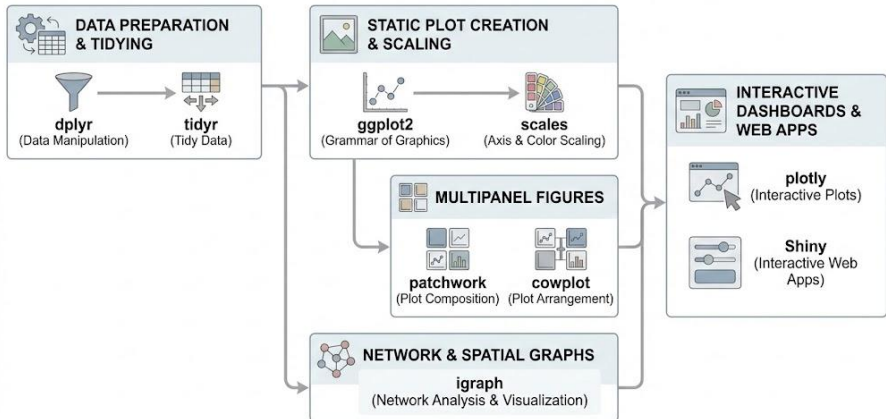


Fig. 8.1 Schematic representation of visualization techniques employed in ecological data analysis using R, ranging from basic statistical plots to advanced multivariate, network, and interactive visualizations.

R PACKAGES FOR ECOLOGICAL DATA VISUALIZATION FLOW



1. GGPLOT2 – THE BACKBONE OF PLOTTING IN R/THE PAINTER’S PALETTE

The **ggplot2** package is the workhorse of data visualization in R. Developed by Hadley Wickham, it implements the **Grammar of Graphics**, which allows you to build plots by layering data, aesthetics (like x/y axes), geometric objects (bars, points, lines), and themes. When it comes to plotting, ggplot2 is your magic wand. It’s like having a paintbrush that follows the grammar of graphics: you layer data, aesthetics, shapes, and themes to create a masterpiece.

Use it for:

- Bar plots (e.g., TDI across sites)
- Boxplots (e.g., guild or trait distributions)
- Line graphs (e.g., spatial or temporal trends)
- Faceted plots (e.g., multi-index comparisons)

Install with:

```
install.packages("ggplot2")
library(ggplot2)
```

2. DPLYR AND TIDYR – DATA WRANGLING ESSENTIALS/ YOUR DATA SCULPTORS

dplyr is used for filtering, summarizing, and transforming data, while **tidyr** helps reshape your data into formats suitable for plotting. Since most of the visualizations rely on “**tidy**” data (where each row is an observation and each column is a variable), these packages are indispensable. Before you even start plotting, you need your data in order. dplyr is like a sculptor’s chisel, it helps you slice, dice, and reshape your dataset. tidyr complements it by turning your data into the “tidy” format that ggplot2 loves: one observation per row, one variable per column.

Use them for:

- Preparing Diathor output for plotting
- Creating long-format data for ggplot2
- Calculating means, medians, and other stats per site/index

Install with:

```
install.packages("dplyr")
install.packages("tidyr")
library(dplyr)
library(tidyr)
```

3. VEGAN – MULTIVARIATE AND ORDINATION PLOTS/ THE ECOLOGIST’S COMPASS

The **vegan** package is the gold standard for **community ecology**, especially for **ordination** techniques such as PCA, NMDS, RDA, and CCA. It allows you to visualize species-environment relationships and explore gradients in complex datasets. When your data has dozens of species and environmental variables, multivariate analysis is key. That's where vegan shines. It's a toolkit for ordination methods like PCA, NMDS, RDA, or CCA methods that reduce complexity and reveal hidden patterns in communities.

Use it for:

- Principal Component Analysis (PCA)
- Redundancy Analysis (RDA)
- Non-Metric Multidimensional Scaling (NMDS)

Install with:

```
install.packages("vegan")
library(vegan)
```

4. PHEATMAP AND COMPLEX HEATMAP – FOR HEATMAPS/ THE PATTERN DETECTIVES

These packages provide beautifully formatted **heatmaps**, which are useful when you want to see patterns across indices, species, or traits at multiple sites. When you want to see patterns at a glance, heatmaps are your friends. pheatmap is simple and effective, while Complex Heatmap lets you get fancy with annotations, clustering, and multiple layers.

Use them for:

- Visualizing index intensity per sample
- Trait clustering across environmental gradients

Install with:

```
install.packages("pheatmap")
library(pheatmap)
```

For advanced customization:

```
BiocManager::install("ComplexHeatmap")
library(ComplexHeatmap)
```

5. FMSB – FOR RADAR (SPIDER) PLOTS

The **fmsb** package is built specifically to generate **radar plots**, which are excellent for comparing multivariate profiles across samples or indices. These plots are circular and intuitive for visualizing performance across multiple axes. Ever wanted to see multiple indices at once, radiating like the spokes of a wheel? fmsb creates radar plots, ideal for comparing sites across multiple axes in a single glance.

Use it for:

- Comparing multiple indices per site in a star-like format

Install with:

```
install.packages("fmsb")
library(fmsb)
```

6 • CORRPLOT – CORRELATION MATRICES/ READING RELATIONSHIPS

To visually explore the **relationships between ecological indices**, the **corrplot** package creates colorful correlation matrices. This is especially useful for examining whether multiple indices show redundancy or complementarity. When you have multiple indices or traits, understanding how they interact is crucial. corrplot turns correlation matrices into colorful, easy-to-read visuals.

Use it for:

- Visualizing pairwise correlations between indices or traits

Install with:

```
install.packages("corrplot")  
library(corrplot)
```

7. LEAFLET – MAPPING DIATOM INDICES SPATIALLY/ MAPS THAT MOVE

The **leaflet** package allows you to create **interactive maps** using latitude and longitude data. This is powerful for visualizing **spatial patterns in water quality**, especially when combined with index values as color gradients or marker sizes. Static maps are informative, but interactive maps are engaging. leaflet lets you plot your sites on actual rivers and lakes, with markers that can vary in color or size according to index values.

Use it for:

- Mapping IPS, TDI, or guild scores across rivers and lakes

Install with:

```
install.packages("leaflet")  
library(leaflet)
```

8. IGRAPH – NETWORK VISUALIZATIONS/ NETWORKS OF LIFE

If you're exploring **interactions or co-occurrences**—such as between taxa and traits, or between sites and environmental parameters—the **igraph** package lets you create network plots and graph theory-based visualizations. Ecology is all about relationships. igraph creates network diagrams, turning species-trait interactions into connected webs.

Use it for:

- Showing species-trait interactions
- Visualizing shared guilds across samples

Install with:

```
install.packages("igraph")  
library(igraph)
```

9. WORDCLOUD – FREQUENCY-BASED VISUALS/ SPOTLIGHT ON DOMINANT SPECIES

The **wordcloud** package lets you turn species frequency data into engaging word clouds. Words are scaled by abundance or frequency, helping highlight dominant taxa in a given dataset. Sometimes, you just want to highlight the most abundant species. wordcloud scales species names by frequency or abundance.

Use it for:

- Representing dominant diatom taxa by abundance

Install with:

```
install.packages("wordcloud")  
library(wordcloud)
```

10. TREEMAP – HIERARCHICAL ABUNDANCE PLOTS/ HIERARCHIES IN A GLANCE

For data with **nested categories** like taxonomic rank or ecological guilds, **treemap** helps represent the hierarchical structure of abundance. Hierarchical data, like taxonomic ranks or guild compositions, shine with treemap. Each block represents abundance, and nested blocks show relationships.

Use it for:

- Visualizing taxonomic dominance
- Guild composition breakdown

Install with:

```
install.packages("treemap")  
library(treemap)
```

11. SHINY – FOR INTERACTIVE DASHBOARDS/ MAKE YOUR DATA DANCE

The **shiny** package is what you need when you want to go from static graphs to **interactive, user-friendly dashboards**. You can embed plots, maps, and data tables in an app format with drop-down filters and sliders. Want interactive dashboards? shiny transforms your static plots into apps with sliders, filters, and dynamic visuals. Your colleagues can explore indices or traits without touching the code.

Use it for:

- Building apps that allow users to explore indices dynamically

Install with:

```
install.packages("shiny")  
library(shiny)
```

12. NETWORK GRAPHS

Network graphs are an excellent way to visualize relationships and interactions in ecological datasets. Using the **igraph** package, network graphs depict entities (nodes) and their connections (edges) in a format that highlights patterns such as **trait similarity** or **taxa interactions**.

For example, diatom species can be represented as nodes, and shared traits like motility or pollution tolerance can form the edges between them. This allows researchers to see which traits cluster together or how species interact in different ecological contexts. Similarly, co-occurrence patterns across sampling sites can be visualized, offering insights into community structure and niche overlap.

Use it for:

- Visualizing **shared traits** or guilds across diatom species.
- Mapping **co-occurrence relationships** between taxa in multiple samples.

Install with:

```
install.packages("igraph")  
library(igraph)
```

13. SUNBURST / CIRCULAR DENDROGRAMS

Sunburst charts and **circular dendrograms** are ideal for visualizing **hierarchical relationships** in data. These visualizations use concentric rings or branching structures to show nested categories, such as taxonomic ranks or trait distributions. Using **plotly**, these charts can also be made interactive, allowing users to explore details by hovering over segments.

In the context of diatom research, a sunburst chart can represent the taxonomic hierarchy of diatoms in a dataset, from phylum to species, while a circular dendrogram can highlight **trait distributions** like size classes or guilds. These plots offer an elegant and intuitive way to explore complex ecological datasets.

Use it for:

- Depicting the **hierarchical structure of taxa**, from genus to species.
- Visualizing **trait composition** and dominance within samples.

Install with:

```
install.packages("plotly")  
library(plotly)
```

14. SPATIAL MAPS

Spatial maps bring ecological data to life by connecting it to geographical locations. With **leaflet**, you can create interactive maps that visualize diatom indices or trait distributions across sampling sites. Each point on the map can represent a site, with markers customized by size or color to reflect ecological metrics like IPS, TDI, or guild proportions.

For example, spatial maps can show how pollution indices change along a river, revealing upstream-to-downstream gradients. They also allow users to identify hotspots of high or low ecological health in a visually engaging way, making them ideal for **monitoring programs** or **decision-making tools**.

Use it for:

- Mapping **IPS, TDI, or guild distributions** across a river basin.
- Visualizing **spatial patterns in ecological indices**.

Install with:

```
install.packages("leaflet")
library(leaflet)
```

15. PCA / ORDINATION PLOTS (FROM VEGAN)

Ordination plots, especially PCA (Principal Component Analysis), RDA (Redundancy Analysis), and NMDS (Non-Metric Multidimensional Scaling), are core tools in community ecology, used to explore ecological gradients, species composition, and environmental relationships. These plots help reduce complex, multidimensional data into 2D or 3D spaces where patterns and clusters become visually interpretable.

In R, the `vegan` package is the go-to tool for producing ordination analyses. Using `rda()` for Redundancy Analysis or `metaMDS()` for NMDS, you can project samples and species onto ordination axes, showing which communities are similar, how they respond to environmental variables, or how species traits correlate with pollution gradients.

When working with DiaThor outputs, PCA or RDA plots can highlight how index scores, trait abundances, or guild distributions change across sites. For example, samples with high TDI and motile guild dominance might cluster separately from those with high IPS and low-profile taxa, revealing underlying pollution gradients or habitat types.

Use it for:

- Exploring community structure in diatom data.
- Visualizing species-environment relationships across sampling sites.

Install with:

```
install.packages("vegan")
library(vegan)
```

16. SANKEY DIAGRAMS

Sankey diagrams are dynamic visual tools that show flows or transitions from one category to another. In ecological analysis, they are especially useful for tracking how species traits, ecological guilds, or functional groups are distributed across different sites or environmental categories. Think of them as river-like flows: wide bands show large groups, while narrower ones show less common transitions. Sankey diagrams are perfect for showing flows: guilds across sites, trait distributions, or species transitions. Width of the bands represents abundance, making patterns instantly visible. Using the `networkD3` package in R, Sankey diagrams can be built from data frames where sources (e.g., taxa or guilds) and targets (e.g., sample sites or seasons) are linked with values (abundance or frequency). These plots elegantly visualize how community structure shifts along environmental gradients or between sampling events.

In DiaThor-based workflows, Sankey diagrams are perfect for showing, for example, how motile, low-profile, and planktonic guilds distribute across polluted vs. clean sites—or how traits like biovolume or chloroplast shape shift seasonally. This kind of visualization provides both quantitative depth and storytelling clarity.

Use it for:

- Visualizing guild shifts across environmental conditions.
- Showing trait distribution patterns between multiple samples.

Install with:

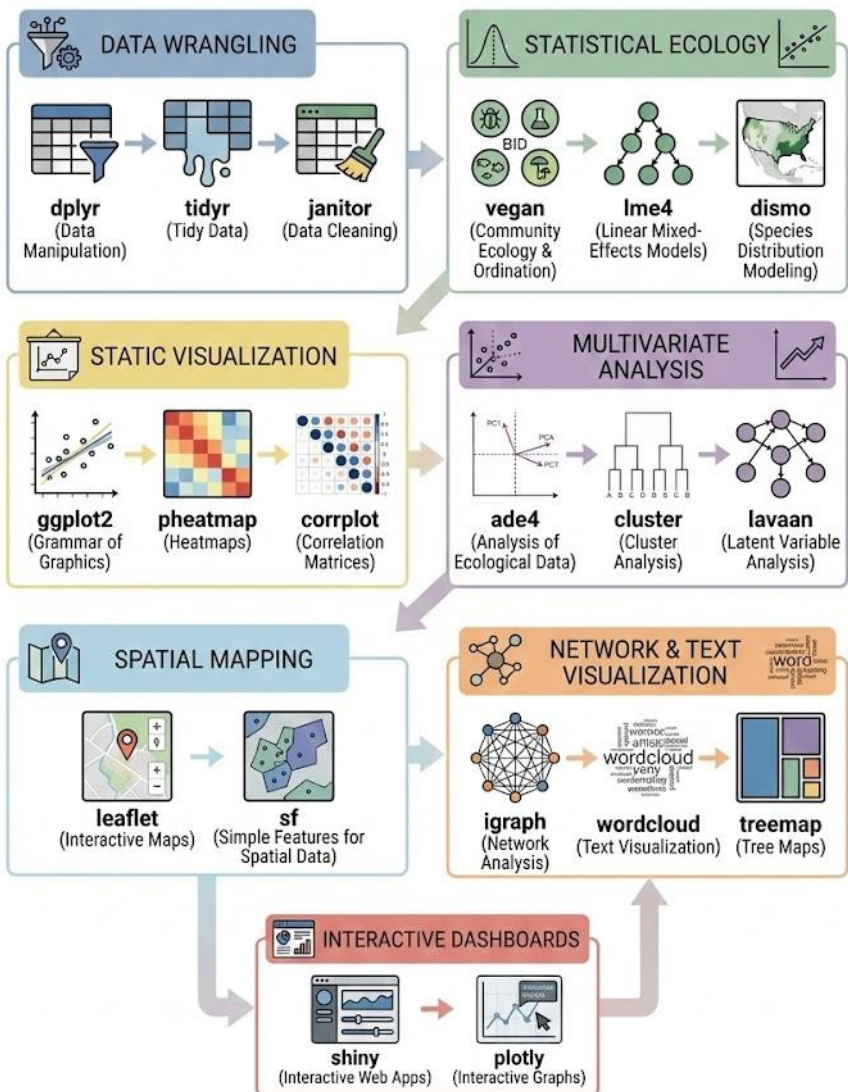
```
install.packages("networkD3")
library(networkD3)
```

17. PLOTLY – SUNBURSTS & CIRCULAR DENDROGRAMS

Hierarchical plots, like sunbursts or circular dendrograms, allow you to explore nested relationships, taxonomic or trait-based. Hover over segments to see details interactively.

```
install.packages("plotly")
library(plotly)
```

ESSENTIAL R PACKAGES FOR ECOLOGICAL DATA ANALYSIS & VISUALIZATION



To get started quickly, you can install all the packages together by giving the command this way:

```
install.packages(c("ggplot2", "dplyr", "tidyr", "pheatmap", "corrplot", "fmsb",  
"leaflet", "igraph", "wordcloud", "treemap", "vegan", "shiny", "plotly"))
```

RECOMMENDED WORKFLOW FOR BEGINNERS

While this chapter introduces a comprehensive suite of visualization tools, beginners should avoid attempting to learn them all at once. We recommend starting with this core workflow:

1. *dplyr & tidyr: Master these first to properly clean, filter, and structure your raw data.*
2. *ggplot2: Use this for all foundational plotting needs (bar charts, scatter plots, line graphs).*

Once you are comfortable with these three core packages, you can gradually explore advanced, specialized tools like `vegan` for ordination or `leaflet` for spatial mapping

CHAPTER 10

A METHODOLOGICAL START: DIATHOR IN PRACTICE

“Mastering a tool begins not in knowing what it does, but in understanding how it thinks. So, when practice meets method, knowledge becomes traceable”

CASE 1: EXECUTING A CLEAN ERROR-FREE FILE

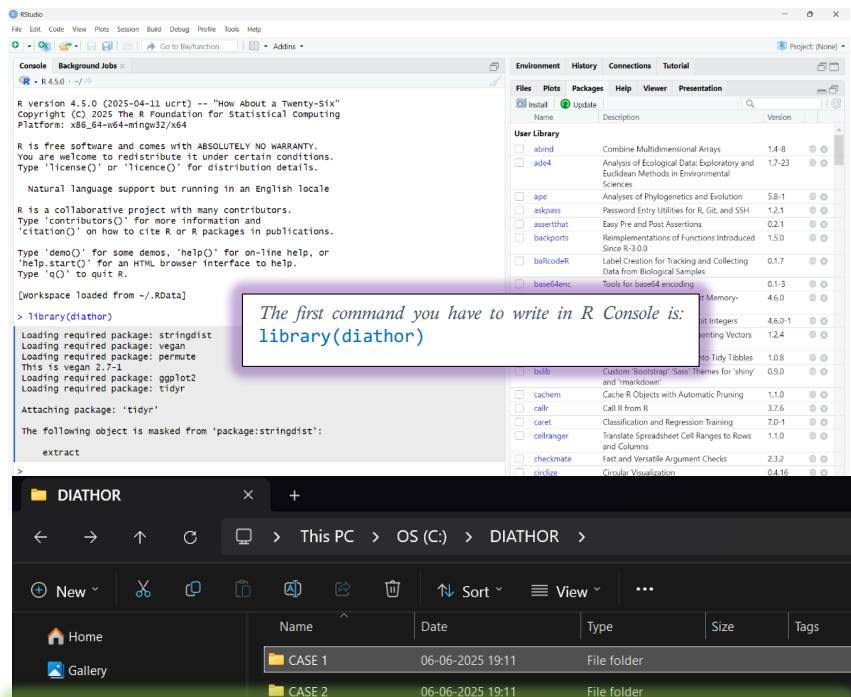
#Step1: Load the package
`library(diathor)`

The first command you must execute in the R Console is to load the DiaThor package into your active environment.

#Step2: Set the Working Directory (Optional but Recommended)

Establish a dedicated folder for your input and output files to maintain an organized workflow.

`setwd("C:/YourFolder/DiatomData")` #Replace with your actual path, if you wish to create a dedicated folder for all your DiaThor related input/output files.



Now you have to select the folder in which you want your output results to get saved. Here, on right clicking the folder you will get a popup, select option "copy as path" or simply select the folder and press `Ctrl+Shift+C` key. Path of the folder will get copied.

The function we have used here is `diaThorAll()`

The moment we will run the command; it will start calculating all the indices collectively and will export the plots in the form of PDF file and results in the form of .exe or .csv file in the folder. Refer images below:

Fig. 9.1 Methodological workflow illustrating the practical application of the DiaThor package using a clean, error-free input dataset.

#Step3: Execute the Master Function

DiaThor is designed primarily as a computational back-end engine to calculate biotic indices rather than an interactive visualization tool. To calculate all available indices simultaneously, use the diaThorAll() master function.

You must define the directory where you want the output files to be saved. Paste your target folder path between the quotation marks. Once executed, a file picker dialog box will appear prompting you to select your input .csv data file.

```
result <- diaThorAll(resultsPath="C:/Users/dell/Desktop/DAITHOR CSV FILES/SAMPLERUN") ←  
Replace with your actual path
```

CRITICAL PATH FORMATTING NOTE:

R Studio requires forward slashes (/) or double backslashes (\\) in file paths. A single backslash (\) will return an error.

Developing a Methodological Mindset

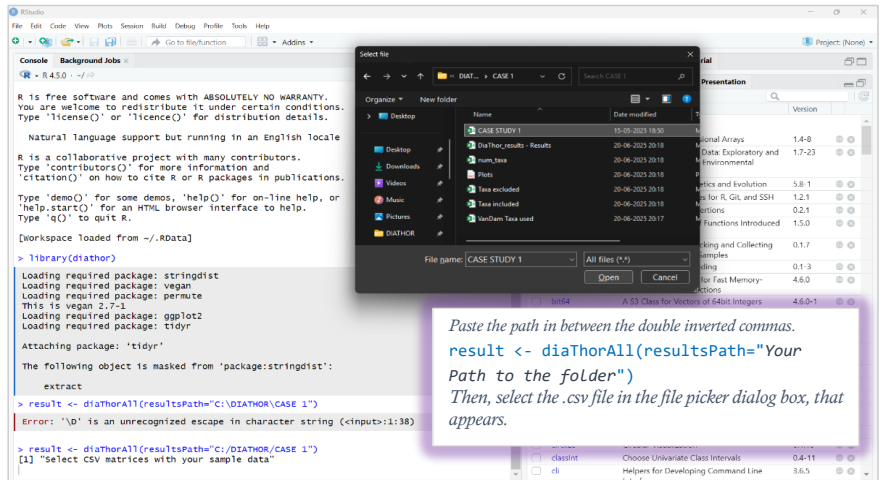
Encountering red error messages during initial software execution is a standard part of the computational learning curve. DiaThor enforces strict data handling rules; it does not merely calculate indices, but rather forces the user to adopt the rigorous data structure required in modern diatom ecology. To operate the package successfully, users must adopt a methodological mindset. DiaThor itself doesn't come with interactive plotting capabilities. It's designed as a back-end engine to calculate biotic indices and metrics, not to dazzle with visuals.

Indices calculated under *diaThorAll()* Function

Function Name	Index / Analysis	Environmental Focus	Purpose / What It Indicates
diat_guilds()	Ecological Guilds	Habitat structure, disturbance	Guild-based ecological strategies
diat_morpho()	Morphological Traits	Functional structure	Cell shape and structure (low
diat_size()	Size Classes	Functional size-based traits	Reflects energy flow and
diat_diversity()	Alpha Diversity	Biodiversity & community	Shannon, Simpson, Evenness indices
diat_vandam()	Van Dam Classification	Saprobity, salinity, trophic state	Environmental preference grouping
diat_disp()	Dispersion	Beta diversity	Variation among sites, heterogeneity

diat_tdi()	Trophic Diatom Index (TDI)	Nutrient enrichment (eutrophication)	Assesses phosphorus pollution and nutrient levels
diat_ips()	Pollution Sensitivity Index (IPS)	Organic/general pollution	Evaluates species sensitivity to general pollution

diat_sla()	Sládeček Index	Organic load (saprobity)	Detects organic pollution from wastewater or decay
diat_idch()	Chilean Diatom Index (IDCH)	General pollution (Chile-specific)	Assesses river pollution in Chilean water bodies
diat_idp()	Indice Diatomique Pédologique (IDP)	Soil diatom community	Soil condition & biological activity
diat_des()	DES Index	General ecological status	Overall diatom community health
diat_epid()	EPID Index	Epiphytic diatom condition	Health of aquatic vegetation zones
diat_ilm()	Italian Diatom Index	General pollution (Italy-specific)	Pollution impact in Italian water systems
diat_lobo()	LOBO Index	Brazilian freshwater systems	Pollution status in Brazilian ecosystems
diat_pbidw()	Pamela Biotic Index (Weighted)	Pollution tolerance	A newer biotic index under development
diat_spear()	SPEARherbicides	Herbicide contamination	Tracks herbicide-sensitive taxa
diat_ddi()	Descy Diatom Index	Pollution (Belgian rivers)	General water quality
diat_idap()	Indice Diatomique d'Activité Pédologique	Soil activity index	Land-use and disturbance in soil habitats
diat_eddi()	Ecological Diatom Index (EDI)	Multiple stressors	Composite ecological quality index
diat_pdise()	PDI-SE	Southeastern USA water quality	Nutrient and pollution stress indicator
diat_checkName()	Taxa Name Checker	Taxonomic standardization	Validates and corrects species
diat_getDiatBarcode()	Diat.Barcode Integration	Species trait data	Connects to Diat.Barcode database
diat_taxaList()	Taxa List Output	Taxonomic information	Lists final cleaned taxa names
diat_guilds()	Ecological Guilds	Habitat structure, disturbance	Guild-based ecological strategies
diathorAll()	All Indices Together	Full assessment	Runs all applicable indices and exports results

**NOTE:**

Correct

```
result <- diaThorAll(resultsPath='\"C:/Users/dell/Desktop/DAITHOR FILES/SAMPLERUN\"')  CSV
```

Incorrect

```
result <- diaThorAll(resultsPath='\"C:\Users\dell\Desktop\DAITHOR FILES\SAMPLERUN\"')  CSV
```

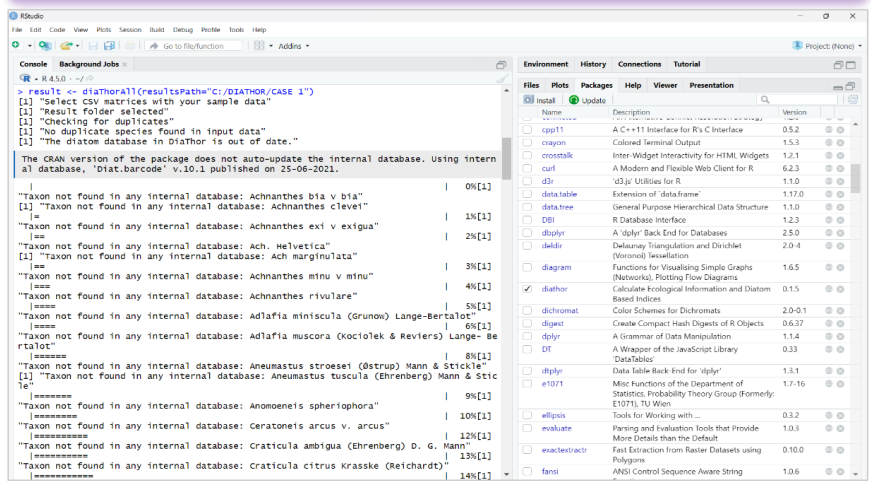


Fig. 9.2 Illustration of Case I, showing DiaThor analysis performed with a correctly formatted and error-free input file.

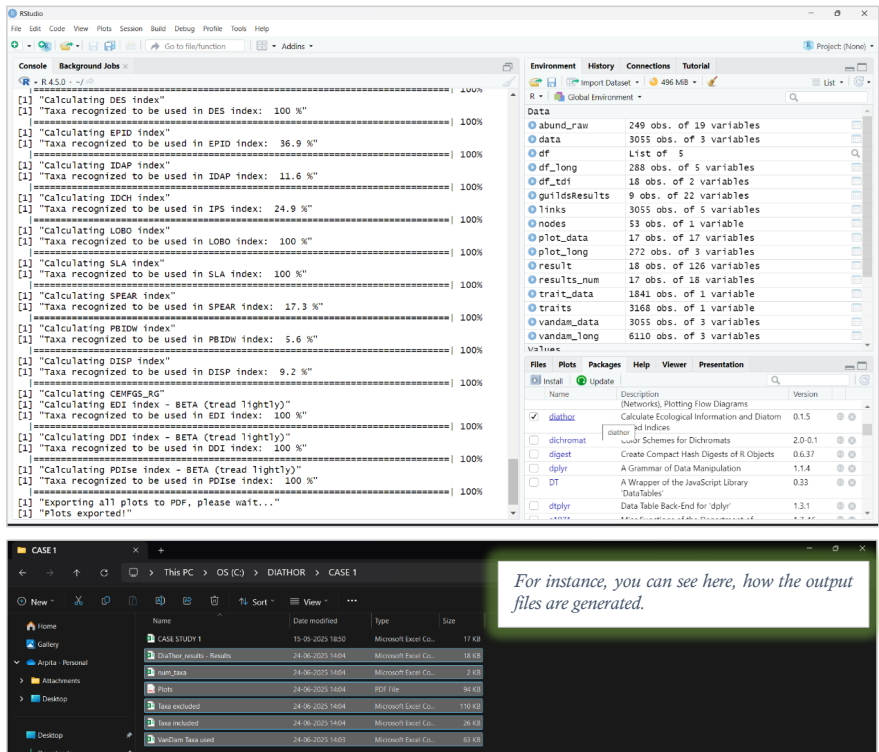


Fig. 9.2 Illustration of Case I, showing DiaThor analysis performed with a correctly formatted and error-free input file.

Step 1: Data Preparation and Standardization

Proper workspace and data preparation in RStudio are critical prerequisites for reliable analysis. Prior to executing the `diaThorAll()` function, users must verify the integrity of their input data. Key considerations include:

Taxonomic Standardization: Ensure the species list is standardized to accepted nomenclature; ambiguous entries (e.g., "Navicula spp.") must be resolved.

Data Format: Confirm that all abundance values are strictly numeric, as text entries will cause computational failures.

Data Type: Clearly establish whether the dataset utilizes absolute counts or relative abundances.

Systematic data preparation is essential for generating accurate ecological insights and preventing analytical errors.

Step 2: Taxonomic Harmonization

The DiaThor package relies on strict taxonomic matching with its internal reference databases. It cannot process localized or non-standard species designations. For instance, inputting a specific variety such as *Gomphonema parvulum* var. *parvulum* may result in an unrecognized taxon error if the internal database only catalogues the species-level designation (*Gomphonema parvulum*). Therefore, taxonomic harmonization is a fundamental methodological requirement to ensure that all input taxa are correctly recognized and analyzed by the software.

Step 3: Initial Index Verification

Rather than processing the entire dataset simultaneously with `diaThorAll()`, it is advisable to initially compute a single index, such as the Trophic Diatom Index (TDI). This step serves as a diagnostic test to verify that the software correctly parses the input data. A successful numerical output indicates proper formatting, whereas an "NA" result signifies unresolved issues within the species list or data structure that require correction.

Step 4: Systematic Troubleshooting

Computational errors should be utilized as diagnostic feedback to understand and resolve formatting discrepancies. For example, an error stating that a species is not found in the database directly indicates a need to standardize the spelling against accepted nomenclature lists. This iterative troubleshooting process systematically improves dataset quality and enhances the reliability of the subsequent analysis.

Step 5: Workflow Documentation

Maintaining a detailed record of the analytical workflow is essential for scientific reproducibility. Researchers should document the specific version of the Diathor package utilized, any taxonomic corrections applied to the original dataset, and the specific indices calculated. This documentation establishes methodological traceability, ensuring that data modifications can be justified during peer review or subsequent analyses.

Step 6: Scaling the Analysis

Following successful initial verification, users can confidently execute comprehensive functions such as `diaThorAll()`. At this stage, the focus shifts to broader analytical tasks, including visualizing temporal trends (e.g., plotting TDI over time) and comparing multiple indices to assess ecological congruence. A rigorous methodological foundation ensures that the resulting ecological interpretations are robust, reproducible, and suitable for scientific publication.

Diathor is not just an R package. It's a conversation partner, a stern but fair mentor. It demands clean input, gives honest output, and teaches you the art of methodological discipline. Mastering Diathor doesn't mean memorizing commands; it means learning to think like a diatom ecologist, with patience, rigor, and a bit of humor.

So, are you ready to stop wrestling with the tool and start practicing with method? Because once you do, every index you calculate will feel less like a number and more like a story told by the rivers themselves.

UNDERSTANDING DIATHOR_ALL() OUTPUTS (STEP-BY-STEP INTERPRETATION)

When you run:

```
diaThor_all("path/to/your/sample_data.csv")
```

The package automatically performs a complete diatom-based assessment and saves results into a new folder. Inside, you'll find several CSV files and plots — let's decode them one by one.

1. Result_indices.csv (or DiaThor_indices.csv)

This is the core output file containing all calculated indices for each sample.

Column Name	Meaning / Interpretation
SampleID	Unique identifier of each water sample you analyzed.
NumTaxa	Number of diatom taxa (species) present in that sample.
IncludedTaxa	Number of taxa successfully matched with Diat.Barcode (used for index computation).
ExcludedTaxa	Taxa not recognized or missing ecological data (not used in indices). A high number here means your sample has species not yet catalogued in Diat.Barcode.
VandamTaxa	Number of taxa for which Van Dam trait data was found.
IPS, TDI, IDAP, etc.	Values of individual diatom indices.
Index_Quality	Often shows the interpretation class (e.g., "Good," "Moderate," "Poor") based on threshold values.

Interpretation:

High Included Taxa with low Excluded Taxa → strong, reliable index computation.

High Excluded Taxa → partial or uncertain assessment (some taxa unmatched).

Compare indices (e.g., IPS vs. TDI): if both show similar class ("Good"/"Poor"), your sample's ecological condition is consistent.

2. EcologicalTraits.csv

This table summarizes Van Dam ecological classifications for the taxa present in each sample.

Each column expresses the proportion or percentage of taxa belonging to each ecological group.

For example:

SampleID	pH_acidobiontic	pH_neutral	pH_alkalibiontic	Oxygen_aerophilous	...
----------	-----------------	------------	------------------	--------------------	-----

Interpretation:

If a sample shows high % alkali biontic → likely alkaline, bicarbonate-rich waters.

If high % aerophilous → diatoms adapted to well-oxygenated conditions (often riffle zones).

Dominance of polysaprobic taxa → organic pollution or sewage influence.

High trophic state (eutrophic) taxa → nutrient enrichment, agricultural or domestic runoff.

These traits help explain why an index (e.g., TDI or IPS) gave its particular value.

Decoding DiaThor Outputs

IPS – Indice de Polluosensibilité Spécifique

- Developed in France, IPS evaluates organic pollution and general water quality.
- Range: 1 (heavily polluted) → 20 (clean water).
- Classification (approximate):
 - 17–20 → High / Excellent quality
 - 13–17 → Good quality
 - 9–13 → Moderate
 - 5–9 → Poor
 - <5 → Very poor
- Ecological meaning:
 - High IPS = dominance of sensitive, pollution-intolerant diatoms (like *Achnanthes minutissimum*).
 - Low IPS = tolerant, eutrophic species (*Nitzschia palea*, *Navicula gregaria*).

IDAP – Index Diatomique Artois-Picardie

- Each emphasizes a slightly different stressor:
- SLA / DESCY: Saprobity (organic matter degradation).
- EPI-D: Eutrophication and organic loading.
- Bacillariophycean Index (BI): General pollution.
- Cross-compare indices. If IPS and TDI agree ("Good" vs "Good"), confidence is high. Divergence hints at mixed stressors (e.g., nutrient enrichment but low organic load).

TDI – Trophic Diatom Index

- Sensitive to organic pollution and flow disturbance, similar to IPS but regionally calibrated for northern France.
- Lower values indicate degradation.

SLA, DESCY, EPID, EPI-D, etc.

- Focuses on nutrient enrichment (eutrophication).
- Range: 0 (oligotrophic) → 100 (highly eutrophic).
- Classes (approx.):
 - 0–30 → Oligotrophic (clean)
 - 31–50 → Mesotrophic (moderate)
 - 51–70 → Eutrophic (enriched)
 - 70 → Hypertrophic (heavily enriched).
- Interpretation:
 - High TDI = dominance of nutrient-loving species (*Gomphonema parvulum*).
 - Low TDI = clean, low-nutrient assemblages (*Fragilaria crotonensis*).

Fig. 9.4 Understanding Diathor Outputs

3. FunctionalTraits.csv

Describes diatom growth forms, motility, guilds, and ecological strategies.

<i>Trait Type</i>	<i>Possible Values / Classes</i>	<i>What It Indicates</i>
Guild	<i>Low-profile, High-profile, Motile, Planktonic</i>	<i>Community structure & disturbance sensitivity.</i>
Motility	<i>Motile / Non-motile</i>	<i>Motile forms thrive in nutrient-rich, unstable habitats.</i>
Size Class	<i>Small / Medium / Large</i>	<i>Related to light penetration and resource competition.</i>
Nitrogen Uptake	<i>Autotrophic, Heterotrophic</i>	<i>Nutrient regime indicator.</i>

Interpretation:

High motile guilds → unstable sediments, higher organic load or siltation.

Dominance of low-profile guilds → stable, oligotrophic, low disturbance systems.

Large cells → more common in nutrient-poor, stable waters.

4. MorphologicalTraits.csv

Covers morphological and anatomical characteristics derived from Diat.Barcode.

<i>Trait</i>	<i>Possible Values</i>	<i>Ecological Meaning</i>
ChloroplastShape	<i>Discoid, H-shaped, Ribbon, Stellate, etc.</i>	<i>Reflects adaptation to light conditions and nutrient status.</i>
ChloroplastNumber	<i>1, 2, Multiple</i>	<i>Related to photosynthetic efficiency.</i>
ChloroplastPosition	<i>Central, Lateral, Parietal</i>	<i>Indicates spatial adaptation within cell.</i>
CellSize / Biovolume	<i>Numeric (μm^3)</i>	<i>Larger biovolume = oligotrophic tendency.</i>

Interpretation :

Morphological traits help explain functional adaptation rather than direct pollution response.

For instance:

- * H-shaped or ribbon chloroplasts → common in high-light, turbulent zones.
- * Parietal chloroplasts → efficient in low-light, nutrient-rich waters. Chloroplast Number
- * Single chloroplast - common in small, oligotrophic taxa; indicates efficient energy use.
- * Two chloroplasts - typical of mid-sized species; balanced adaptation.
- * Multiple chloroplasts - seen in large, motile, or shade-adapted diatoms; higher photosynthetic flexibility.

Communities with numerous multi-chloroplast taxa often indicate low-light, nutrient-enriched environments.

Chloroplast Position:

The spatial position (central, lateral, apical, or parietal) influences light harvesting and nutrient exchange:

- * Central - efficient for evenly lit conditions.
- * Lateral / Parietal - adapt to grazing pressure or light variability.
- * Apical - seen in elongated forms adapted for streamlining flow or sediment avoidance.

Though subtle, these positional traits are powerful indicators of habitat stability and hydrological stress.

5. Taxa_List.csv or Matched_Taxa.csv

Lists all taxa found in your sample along with whether each was included or excluded in index computation, plus the matched Diat.Barcode ID.

<i>TaxonName</i>	<i>Matched</i>	<i>Diat.BarcodeID</i>	<i>IncludedInIndices</i>	<i>ExcludedReason</i>
------------------	----------------	-----------------------	--------------------------	-----------------------

Interpretation :

Taxa with Matched = FALSE → spelling errors, missing species, or outdated names.

You can correct those names (check Diat.Barcode online) and rerun DiaThor_all() for better accuracy.

6. Plot Outputs (inside “Plots” folder)

DiaThor also creates pre-made plots such as:

Index Comparison Bar plots: show variation of indices (e.g., IPS, TDI) across sites.

Trait Composition Plots: visualize ecological or morphological trait proportions.

Taxa Count Plot: displays total taxa, included, excluded, and Van Dam taxa per sample.

Interpretation :

Look at the balance between included and excluded taxa — this determines confidence in your data.

If excluded taxa form >25% of total, interpret results cautiously (and consider revising taxonomy).

7. Understanding Key Numeric Columns

In many DiaThor outputs, you’ll encounter these core variables:

<i>Variable</i>	<i>What It Represents</i>	<i>How to Use It</i>
<i>numTaxa</i>	Total number of diatom species recorded	Species richness measure
<i>numTaxaIncluded</i>	Count of taxa used in index calculation	Indicates data completeness
<i>numTaxaExcluded</i>	Count of taxa not found in Diat.Barcode	High value = lower confidence
<i>VandamTaxa</i>	Taxa for which Van Dam traits were retrieved	Used in trait-based summaries
<i>NotEvaluatedTaxa</i>	No ecological data available	Usually rare or new species

Example:

If a sample has

numTaxa = 80, numTaxaIncluded = 72, numTaxaExcluded = 8,
you can report:

> “Index reliability was high (90% of taxa included in computation), indicating comprehensive ecological coverage.”

8. Practical Interpretation Summary

<i>Output Component</i>	<i>Meaning</i>	<i>Use for Biologists</i>
<i>Indices (IPS, TDI, etc.)</i>	<i>Numeric water quality rating</i>	<i>Direct pollution evaluation</i>
<i>Van Dam traits</i>	<i>Ecological preference summaries</i>	<i>Explains environmental conditions</i>
<i>Functional traits</i>	<i>Guild and motility structure</i>	<i>Habitat stability & nutrient indication</i>
<i>Morphological traits</i>	<i>Cell and chloroplast diversity</i>	<i>Light and energy adaptation</i>
<i>Taxa inclusion report</i>	<i>Quality assurance</i>	<i>Data cleaning and reliability check</i>

9. Why Some Indices (like DAIPo) Might Be Missing

DAIPo (Diatom Assemblage Index to organic pollution) requires specific ecological coefficients available only for certain regional datasets (mainly France, Italy, or Spain).

If your taxa are not in that reference list, DiaThor automatically skips the index.

So, it's not a software error – just data unavailability. You can verify by checking the ExcludedIndices.csv file (if generated).

10. How to Report Your Results:

The output of DiaThor_all() was analyzed to interpret diatom-based ecological status.

pH Preference

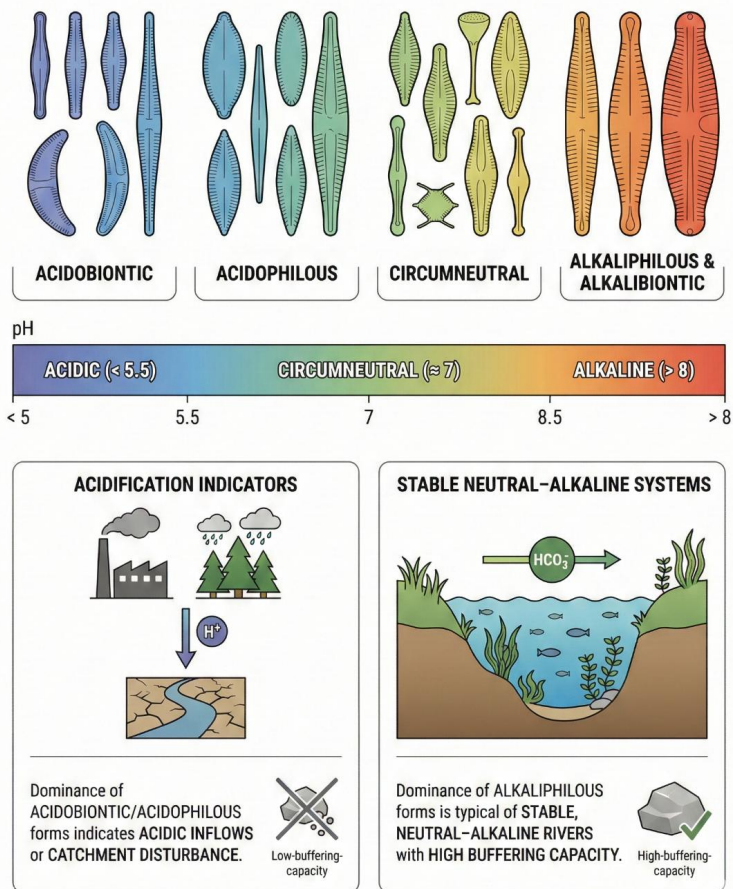
Each taxon is assigned to one of several pH classes:

- Acidobiontic – thrives exclusively in acidic environments (pH < 5).
- Acidophilous – prefers acidic to slightly neutral waters.
- Circumneutral – tolerant of near-neutral pH.
- Alkaliphilous / Alkalibiontic – found in alkaline conditions (pH > 7.5).

This trait helps in detecting acidification or buffering capacity of water bodies.

In neutral to alkaline Indian rivers, dominance of alkaliphilous species is common, while increased acidophilous forms may signal acidic inflows or catchment disturbances.

DIATOM pH PREFERENCE CLASSES & ECOLOGICAL IMPLICATIONS



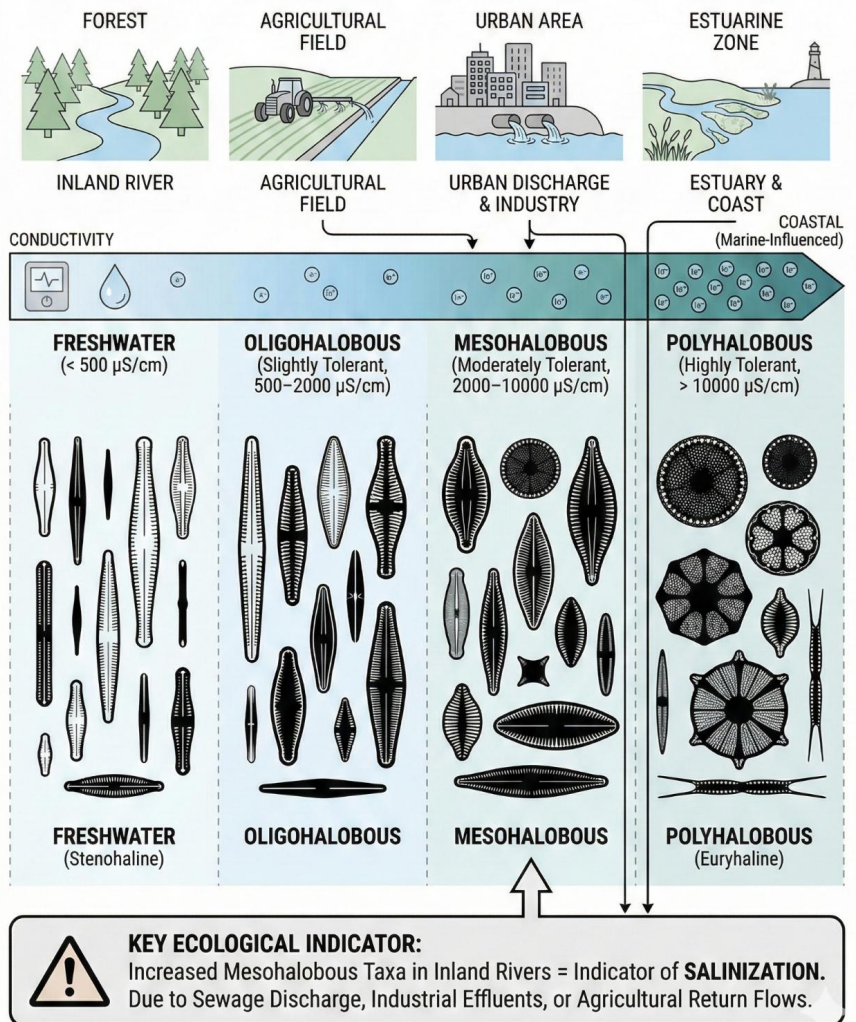
Salinity Tolerance

Diatoms respond sensitively to dissolved salts. Van Dam salinity groups include:

Freshwater – limited to low-conductivity habitats.

- **Oligohalobous** – tolerate mild salinity.
- **Mesohalobous** – thrive in moderately saline waters.

DIATOM SALINITY TOLERANCE GROUPS & ECOLOGICAL IMPLICATIONS



- Polyhalobous – characteristic of estuarine or coastal zones.

In inland rivers like Yamuna or Betwa, an increased share of mesohalobous taxa often reflects salinization from urban sewage, industrial discharges, or agricultural return flows.

Oxygen Requirement

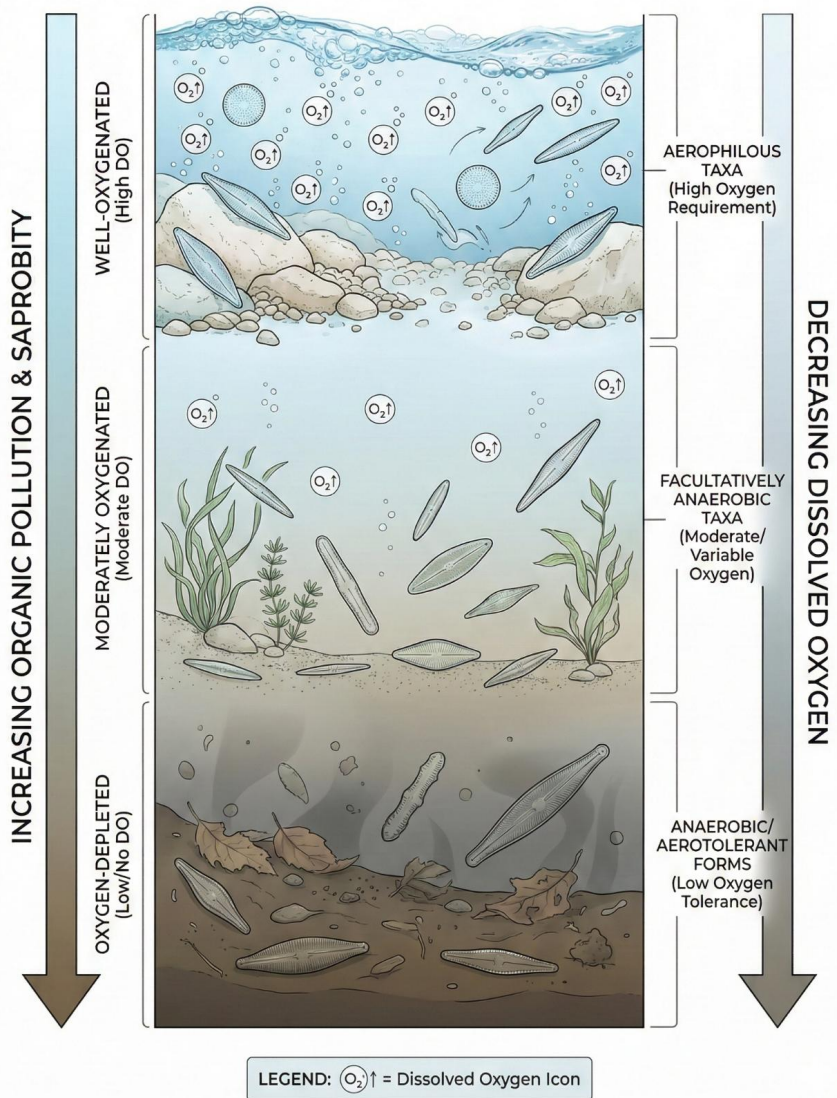
Oxygen preference is a sensitive bioindicator of organic pollution:

Aerophilous species prefer high oxygen (clean, well-mixed water).

Facultatively anaerobic taxa tolerate lower oxygen.

Anaerobic or aerotolerant forms dominate under oxygen-depleted, high organic load conditions.
Hence, oxygen-related traits align strongly with saprobity and trophic indices such as IPS or EPI-D.

DIATOM OXYGEN PREFERENCE TRAITS & ORGANIC POLLUTION



Nitrogen Uptake Metabolism

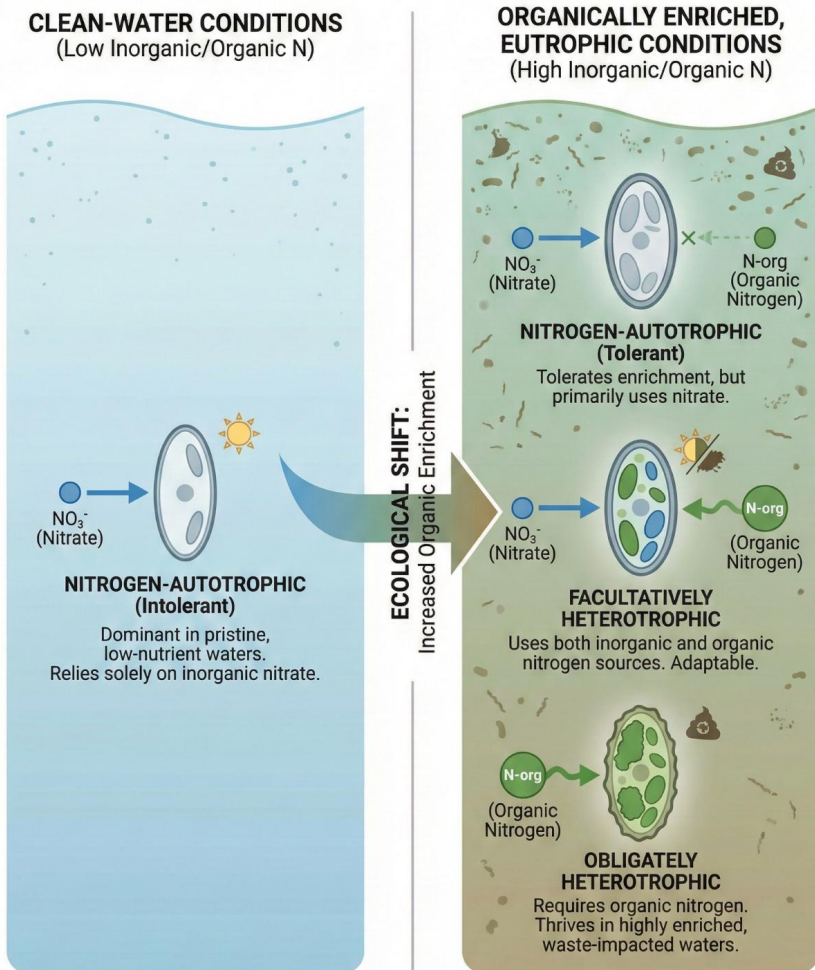
This trait reflects how diatoms utilize nitrogen sources:

- Nitrogen-autotrophic (tolerant) – use nitrate as main N source; dominate in clean water.

- **Nitrogen-autotrophic (intolerant)** – sensitive to organic nitrogen.
- **Facultatively heterotrophic** – can use organic nitrogen.
- **Obligately heterotrophic** – rely entirely on organic forms.

A shift from autotrophic to heterotrophic taxa usually signals organic enrichment or eutrophication.

DIATOM NITROGEN UPTAKE STRATEGIES & ECOLOGICAL SHIFT



ECOLOGICAL INTERPRETATION:

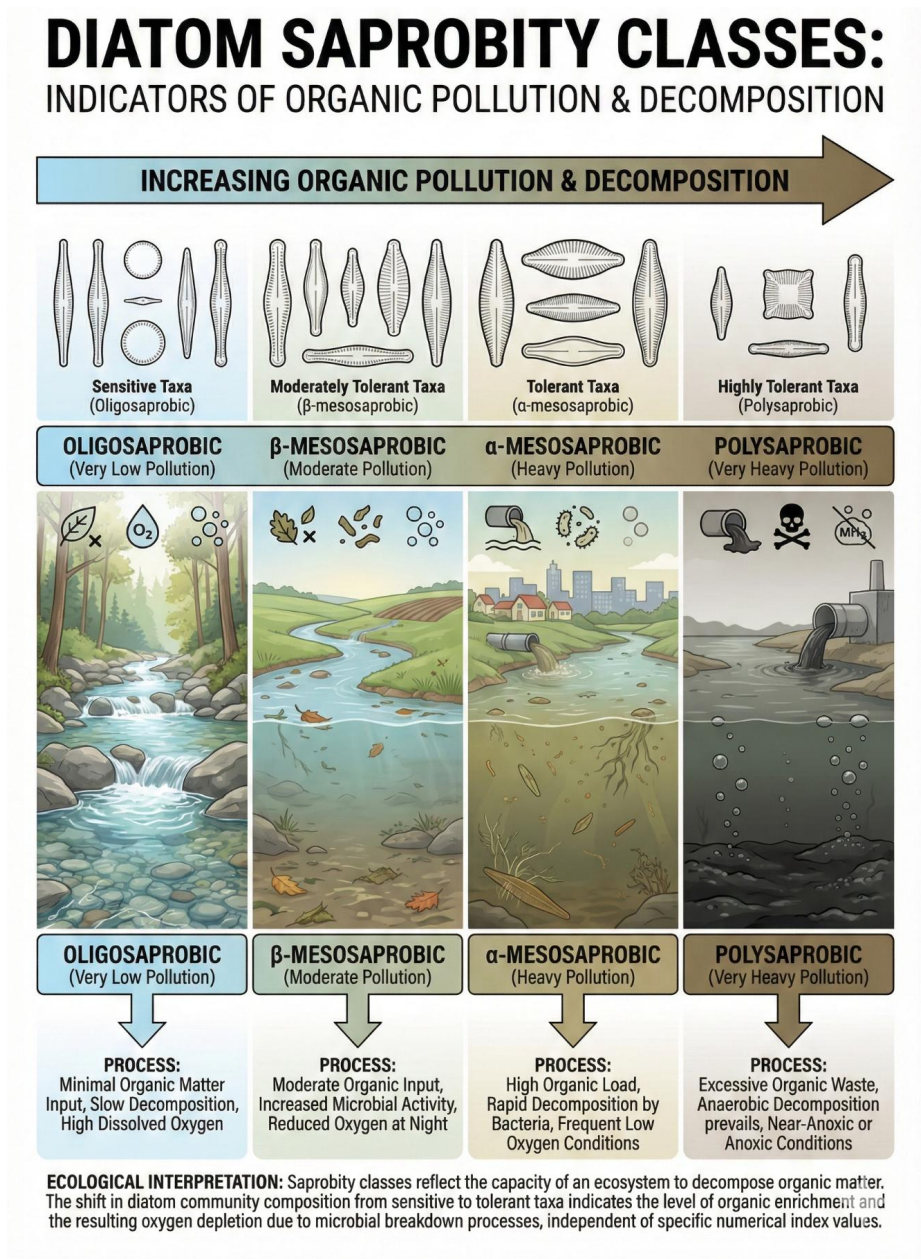
Shift from autotrophic to heterotrophic dominance indicates increasing organic pollution and eutrophication.

Saprobity

Saprobity indicates tolerance to organic matter decomposition:

- **Oligosaprobic** – sensitive, clean water indicators.
- **β -mesosaprobic** – slight pollution tolerance.
- **α -mesosaprobic** – moderately polluted water
- **Polysaprobic** – high organic load, degraded environments.

This classification closely parallels indices like IPS or IDSE but provides ecological explanation rather than just a score.



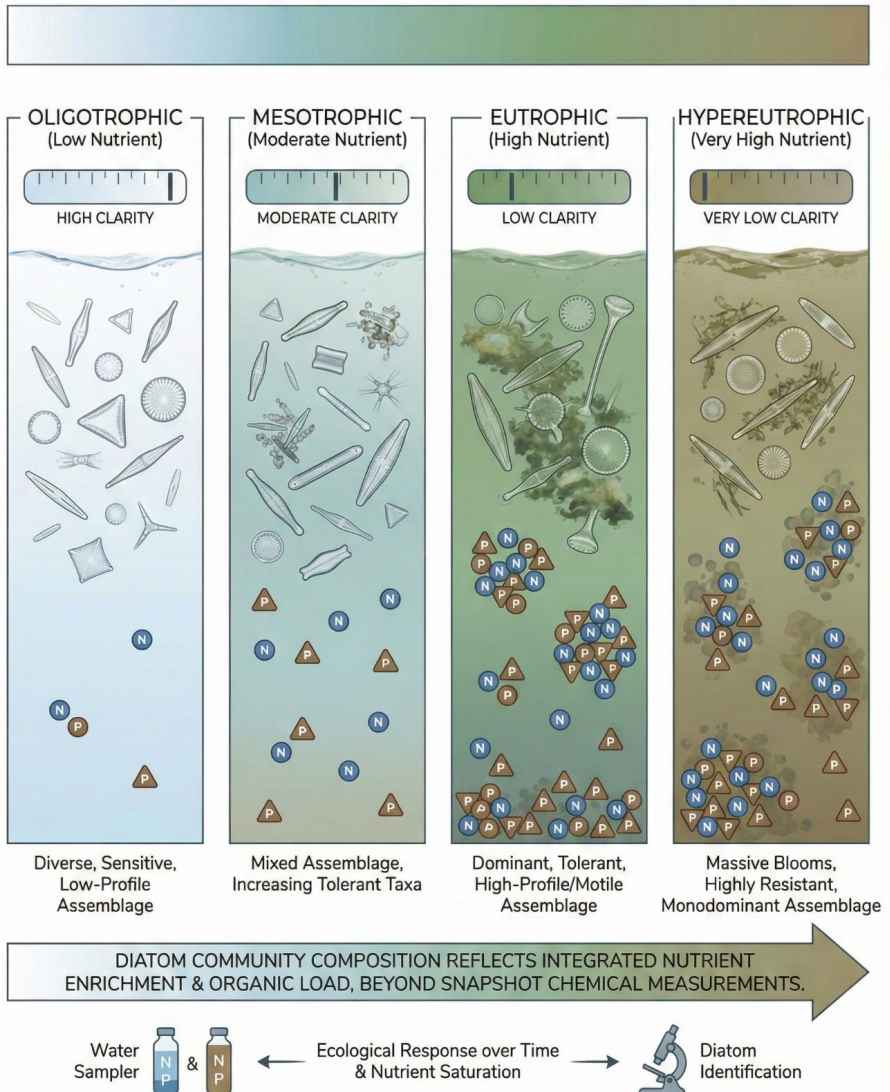
Trophic State

Represents nutrient status preference:

- **Oligotrophic** – nutrient-poor habitats.
- **Mesotrophic** – moderate nutrient availability.
- **Eutrophic** – nutrient-rich systems.

TROPHIC STATE INTERPRETATION USING DIATOMS: ECOLOGICAL GRADIENT & COMMUNITY RESPONSE

ECOLOGICAL GRADIENT: OLIGOTROPHIC TO HYPEREUTROPHIC CONDITIONS



- **Hypereutrophic – very high nutrient and organic loads.**

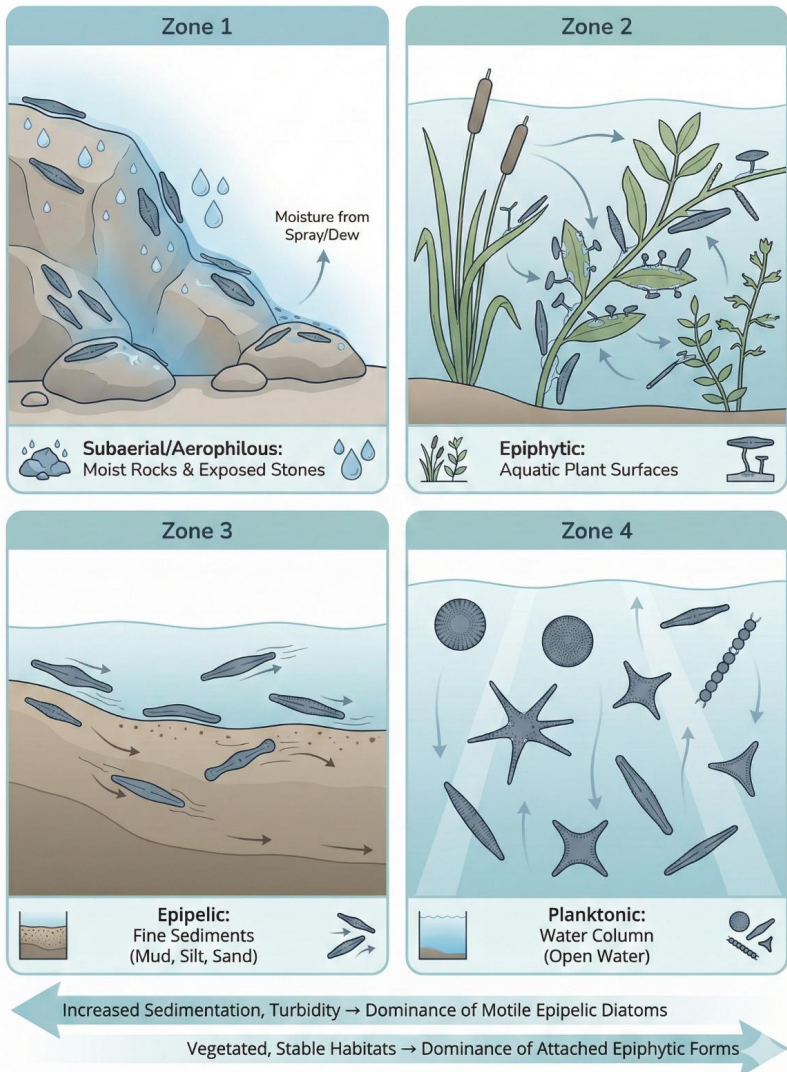
By comparing trophic preference data with the TDI index, users can validate whether the diatom community's structure aligns with the measured nutrient level.

Moisture and Habitat Type

- **Subaerial / Aerophilous** – on moist surfaces, exposed stones, or rocks.
- **Epiphytic** – attached to aquatic plants.
- **Epipelic** – growing on sediments.
- **Planktonic** – freely floating or weakly attached.

Dominance of epipelic and motile diatoms suggests sedimentation or turbidity, whereas epiphytic forms thrive in vegetated, stable habitats.

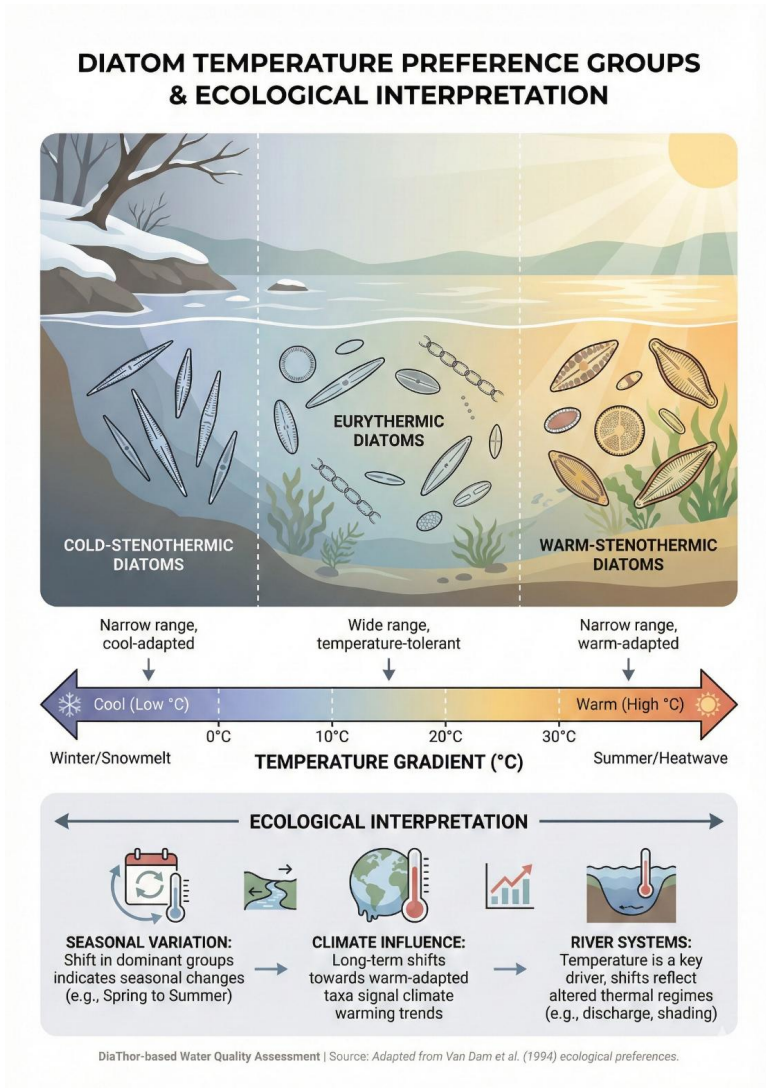
DIATOM MOISTURE & HABITAT TYPES



Temperature Preference

- **Cold-stenothermic** – restricted to cool waters.
- **Warm-stenothermic** – prefer higher temperatures.
- **Eurythermic** – tolerant across wide temperature ranges.

Although temperature preference is less critical in tropical rivers, shifts in these groups can indicate seasonal or climate-related influences.



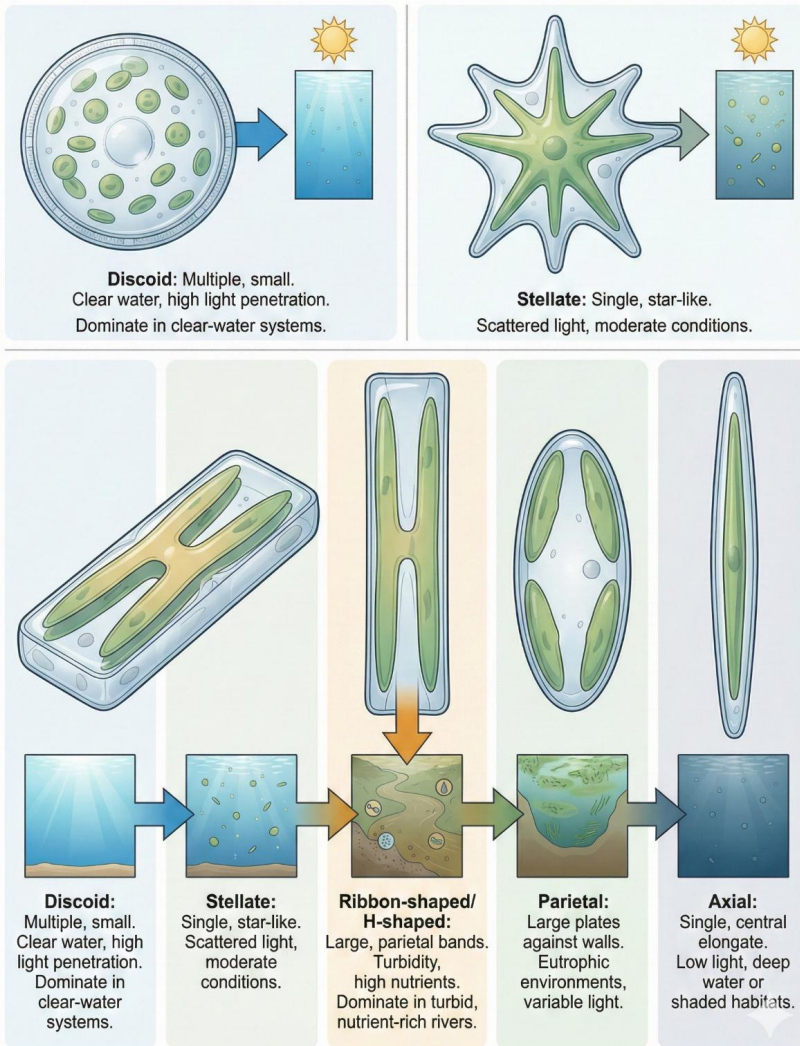
Apart from ecological preferences, DiaThor extracts morphological features that reflect physiological strategies.

Chloroplast Shape

Diatoms exhibit diverse chloroplast morphologies:

- **Discoïd** (e.g., *Cyclotella*): efficient in vertical light columns.
- **Stellate** (e.g., *Achnanthes*): optimal for scattered light.

DIATOM CHLOROPLAST SHAPE DIVERSITY & ENVIRONMENTAL INDICATIONS



- **Ribbon-shaped or H-shaped (e.g., Navicula, Nitzschia):** adaptive for motile taxa and high surface area for photosynthesis.
- **Parietal or Axial:** position-dependent adaptation to turbidity or shading.

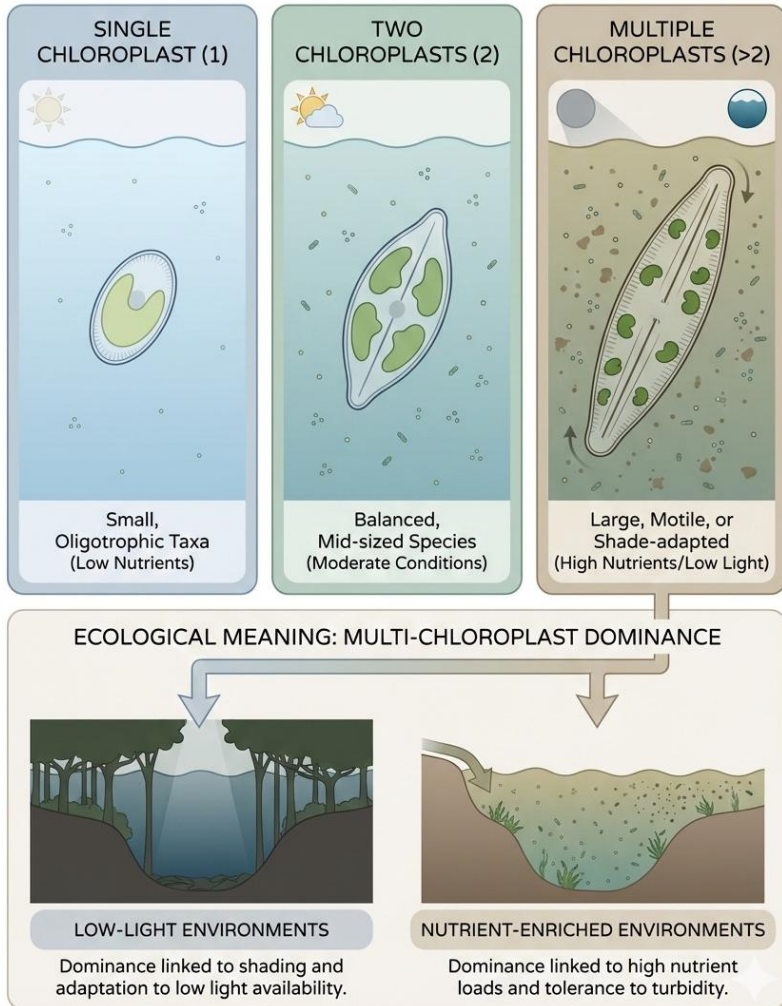
Chloroplast shape diversity offers insights into light adaptation and habitat stability. In turbid or eutrophic rivers, motile ribbon-shaped chloroplast types dominate, while discoid forms often mark clear-water systems.

Chloroplast Number

- **Single chloroplast** – common in small, oligotrophic taxa; indicates efficient energy use.
- **Two chloroplasts** – typical of mid-sized species; balanced adaptation.
- **Multiple chloroplasts** – seen in large, motile, or shade-adapted diatoms; higher photosynthetic flexibility.

Communities with numerous multi-chloroplast taxa often indicate low-light, nutrient-enriched environments.

VARIATION IN CHLOROPLAST NUMBER IN DIATOMS & ECOLOGICAL ASSOCIATIONS



Chloroplast Position

The spatial position (central, lateral, apical, or parietal) influences light harvesting and nutrient exchange:

- **Central** – efficient for evenly lit conditions.
- **Lateral / Parietal** – adapt to grazing pressure or light variability.
- **Apical** – seen in elongated forms adapted for streamlining flow or sediment avoidance.

Though subtle, these positional traits are powerful indicators of habitat stability and hydrological stress.

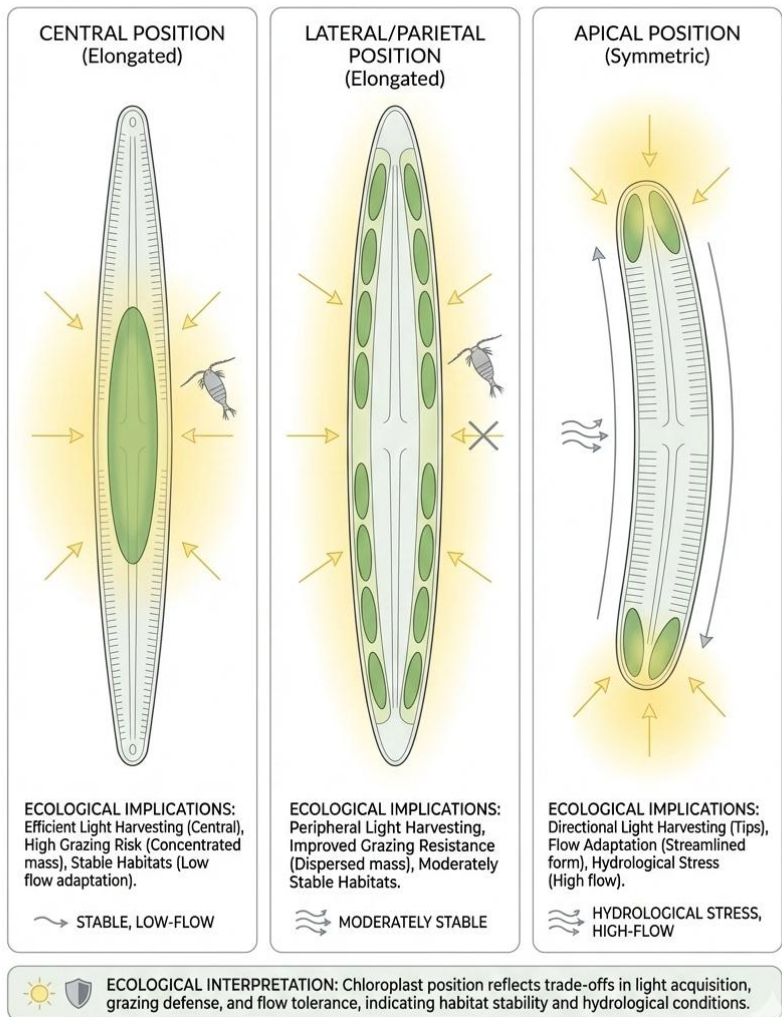
Cell Size, Shape, and Biovolume

DiaThor calculates average cell biovolume, size class, and length–width ratio for each taxon.

- Small-celled species (low biovolume) dominate under nutrient-rich or disturbed conditions.
- Large-celled diatoms thrive in stable, low-nutrient, low-disturbance habitats.

Hence, a reduction in average biovolume or dominance of nano-sized taxa may reflect ecological degradation.

DIATOM CHLOROPLAST POSITION & ECOLOGICAL IMPLICATIONS

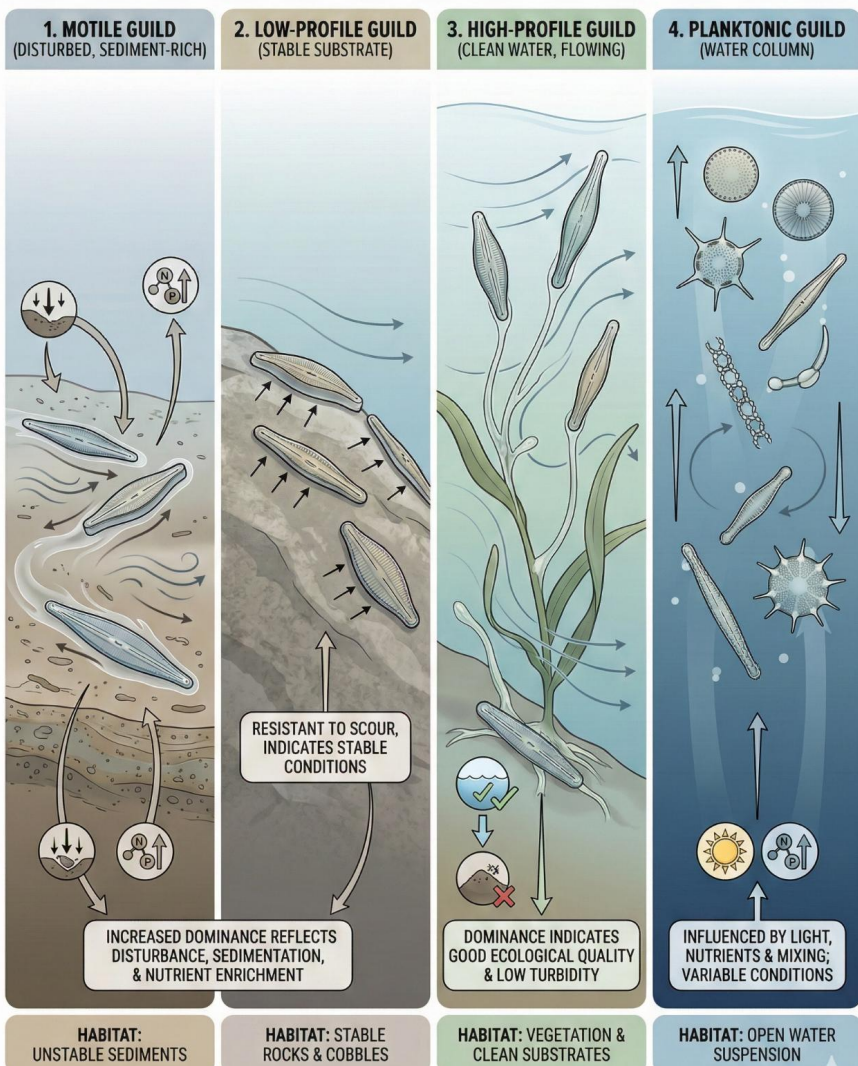


Motility

Diatoms are classified as:

- Motile – capable of gliding; adapt to fine sediments and nutrient pulses.
- Low-profile (attached) – tolerate low disturbance; common in clean, stable substrata.
- High-profile (stalked) – sensitive to turbidity, indicators of high-quality water.

DIATOM MOTILITY & ECOLOGICAL GUILDS: INDICATORS OF ENVIRONMENTAL CONDITION



Arrows/Flow Lines: Movement/Flow/Ecological Link
Icons: Sediment, Nutrients, Water Quality
Diatom Shapes: Simplified Representations

- **Planktonic – suspended in water column; often respond to eutrophication.**

This trait is extremely informative:

An increase in motile guild proportion indicates instability, sedimentation, or enrichment, while high-profile forms suggest good ecological conditions.

Trait-based outputs complement numerical indices by providing:

Mechanistic interpretation of pollution response.

Insights into functional redundancy or loss within diatom communities.

Indications of multiple stressors (light, sediment, salinity, organic matter).

For example:

- A site with low IPS but dominated by motile, ribbon-chloroplast, heterotrophic taxa clearly signals organic enrichment and turbidity.
- Conversely, dominance of high-profile, alkaliphilous, and aerophilous species indicates stable, oxygen-rich, oligotrophic conditions.
- Each sample's index scores (IPS, TDI, etc.) were cross-examined with Van Dam trait distributions and morphological attributes.
- A low number of excluded taxa (<10%) confirmed strong taxonomic coverage.
- Dominance of alkali biontic and motile diatoms, coupled with high TDI values, indicated eutrophic and moderately disturbed conditions.
- Conversely, samples with higher proportion of low-profile guilds and single-disc chloroplast shapes corresponded to oligotrophic, clear-water habitats.
- * Dominance of alkali biontic + eutrophic + aerophilous taxa → alkaline, oxygen-rich but nutrient-enriched systems (agricultural landscapes).
- * Polysaprobic + anaerobic dominance → stagnant, polluted, possibly sewage-impacted waters.
- * Acidobiontic + oligosaprobic → highland, low-impact forest streams.
- * Van Dam traits “translate” your community structure into a habitat profile. If indices say “Good,” traits tell why

The beauty of DiaThor lies in its synthesis, blending numerical precision with ecological storytelling.

Each file is a chapter in the life of your river: indices reveal health, traits reveal personality, and morphology reveals resilience.

Together, they let you hear the whispers of diatoms, the sentinels of aquatic truth.

Trait-based assessment within DiaThor enriches the understanding of diatom communities far beyond pollution indices.

It connects biological form and function to environmental gradients, allowing biologists, even without coding expertise to grasp why certain diatom assemblages thrive under specific conditions.

By integrating these traits into routine analysis, researchers can:

- *Link diatom structure with ecological processes.*
- *Support water quality indices with biological explanations.*
- *Build stronger arguments in ecological and environmental reporting.*

DiaThor thus serves not only as a computational tool but as an ecological interpreter, translating community structure into the language of environmental meaning.

CHAPTER 11

PLOTTING AND VISUALIZATION

“Science becomes insight when data is made visible and relationships revealed”

In the modern era of data-driven ecology, the ability to visualize data is as critical as generating it. While raw numerical data contains the foundational details of an ecosystem, plotting and visualization make these details accessible, enabling the recognition of patterns and the generation of insights. For ecological datasets—especially those involving multiple indices, taxa, and sampling sites—visual representation is essential for interpreting trends, comparing sampling locations, and effectively communicating results to both scientific and non-scientific audiences.

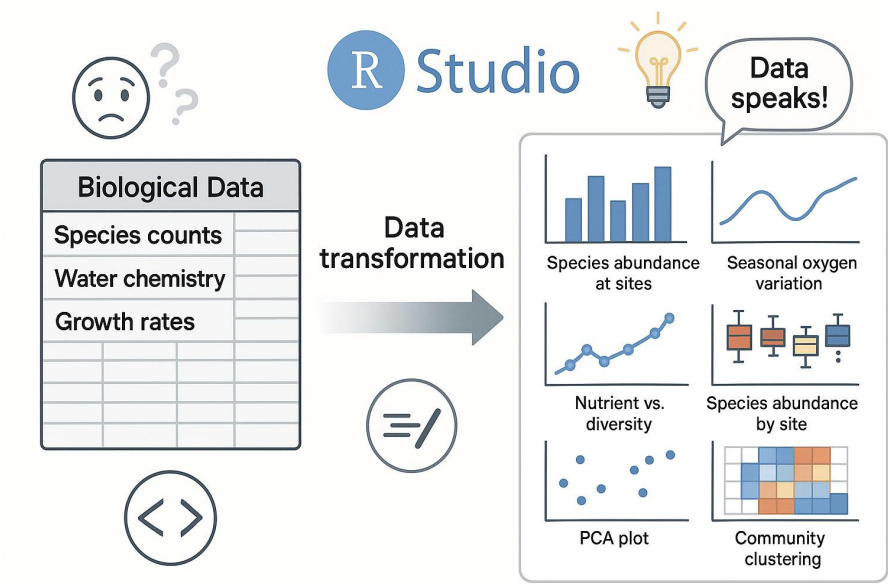
Leveraging RStudio for Reproducible Graphics

RStudio, the integrated development environment (IDE) for R, has revolutionized scientific visualization by offering robust, customizable plotting tools that combine statistical accuracy with publication-ready aesthetics. Utilizing libraries such as ggplot2, lattice, and plotly, researchers can construct plots that are both visually compelling and deeply informative.

Key Benefits of Plotting Data:

- **Detecting Patterns:** Visualization enables researchers to rapidly detect spatial gradients, seasonal trends, and pollutant impacts.
- **Identifying Errors:** Anomalies or data quality issues become immediately apparent in graphical formats.
- **Effective Communication:** A well-designed graph captures relationships that may be missed in tabular data, enhancing communication with environmental managers and policymakers.

A key advantage of R is its ability to layer complex information without cluttering the visual output. For instance, a basic bar chart of diatom indices across various sites can be systematically enhanced with color-coded categories, error bars, facet grids, and interactive tooltips. This transforms a simple chart into a comprehensive, data-driven narrative. Furthermore, visualization in R is inherently reproducible; the same script can be executed with new or updated datasets, ensuring methodological consistency across different projects and time periods.



Visualization in the Context of Diathor

Within the context of diatom-based water quality assessment, the Diathor package exemplifies how ecological data can be seamlessly processed and visualized within the R environment. After computing a comprehensive suite of

biotic indices and ecological traits using the `diaThorAll()` function, users generate a rich dataset that achieves maximum clarity through effective visualization.

Graphical Capabilities of DiaThor

DiaThor offers versatile plotting capabilities, generating visualizations either through its built-in `plotAll = TRUE` argument or via external integration with packages like `ggplot2`. These capabilities allow users to generate:

- **Bar plots illustrating index values across multiple sampling sites.**
- **Boxplots displaying variations in chloroplast shape, size classes, or biovolume.**
- **Stacked bar charts depicting guild composition (e.g., motile, low-profile, planktonic forms).**
- **Facet plots for side-by-side comparisons of multiple ecological indices.**
- **Line graphs designed to examine spatial or temporal gradients.**

Translating Indices into Insights

These plots do more than display raw data; they actively translate ecological relationships. For example, a simultaneous decrease in the Specific Pollution Sensitivity Index (IPS) and an increase in the Trophic Diatom Index (TDI) along a polluted stream gradient becomes visually distinct. Similarly, a plotted shift in guild structure from high-profile to motile forms clearly illustrates increasing sediment stress. Because DiaThor exports structured, tidy CSV files, users can fully exploit R's broader visualization ecosystem. By integrating DiaThor outputs with tools like `vegan` for ordination, `plotly` for interactivity, or `leaflet` for spatial mapping, diatom data can be incorporated into multi-layered, dynamic dashboards ideal for reports, presentations, and policy communication.

The Power of Visualization in Ecological Insight

When a graph reveals that a historically pristine stream is trending toward eutrophication, or that a polar habitat is undergoing warming based on shifts in species traits, it transcends basic illustration to become scientific evidence and a catalyst for action. For researchers, students, and environmental managers, plotting diatom indices is a critical analytical step. Raw numerical data—whether sourced from DiaThor, field notebooks, or external spreadsheets—requires visualization to communicate its underlying ecological meaning effectively.

Step 1: The Rationale for Visualization

Consider the challenge of explaining diatom indices to policymakers using raw values alone (e.g., Site A: TDI = 72.3; Site B: TDI = 41.6; Site C: TDI = 28.9). Presenting these figures as a clear bar plot instantly communicates the environmental reality: Site A is highly eutrophic, Site C is comparatively stable, and Site B represents a transitional state. This demonstrates the immediate communicative power of graphical data representation.

Step 2: Defining the Visual Objective

Before generating a plot, researchers must identify the primary pattern they intend to highlight within their dataset (e.g., dissolved oxygen trends, pH variations, or IPS values). Determining whether the focus is on site-to-site differences, seasonal temporal trends, or correlations with land use will dictate the most appropriate visualization method.

Step 3: The R Visualization Toolbox

RStudio offers a comprehensive suite of visualization packages tailored for specific analytical needs:

- **ggplot2:** The foundational tool for creating elegant, highly customizable static plots.
- **plotly:** Utilized for generating interactive plots featuring hover text, zooming, and clicking capabilities.
- **sf / tmap / leaflet:** Essential for spatial analysis when maps communicate spatial distributions more effectively than tables.
- **vegan:** The standard for generating multivariate ordination plots, including NMDS, PCA, and CCA.
- **radarchart / fmsb:** Ideal for creating multi-index radar plots to compare performance across several axes simultaneously.
- **ComplexHeatmap / pheatmap:** Deployed to uncover clustering and hierarchical patterns within large, multidimensional datasets.
- **circlize / chorddiag:** Used for illustrating complex relationships, directional flows, and network interactions.

Visualization is a translation process that converts raw data into comprehensible scientific narratives. The primary objective is to illuminate data with clarity rather than obscure it with unnecessary graphical complexity. Ultimately, transforming computed DiaThor indices into well-structured visualizations is the critical step where data becomes actionable ecological insight.

- Now, you show them this simple bar plot: *#Refer Fig. 10.1*

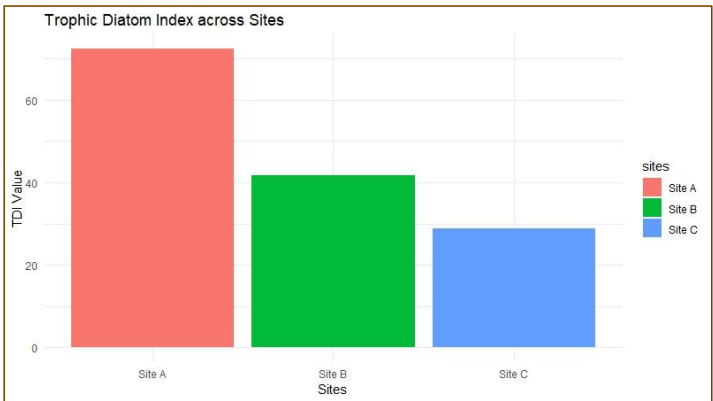


Fig. 10.1 Comparison of Trophic Diatom Index (TDI) values across sampling sites, illustrating spatial variation in nutrient enrichment conditions.

Suddenly, they get it: Site A is eutrophic, Site C is struggling, and Site B is in the middle. That's the power of pictures.

Pause here. Think of your own dataset, maybe dissolved oxygen, maybe pH, maybe IPS values.

What's one pattern you want people to see immediately?

Is it a difference among sites? A seasonal trend? Or a correlation with land use?

Hold that thought, because that vision guides your choice of plot.

In RStudio, you're spoiled for choice. Here are your artist's brushes:

- **ggplot2** → your Swiss Army knife, elegant for almost everything. *#Refer Fig. 10.2*



Fig. 10.2 Seasonal variation in Trophic Diatom Index (TDI) values across sampling months.

- **plotly** → when you want interactivity (hover, zoom, click). *#Refer Fig. 10.3*
- **sf / tmap / leaflet** → when maps tell the story better than tables. *#Refer Fig. 10.4*

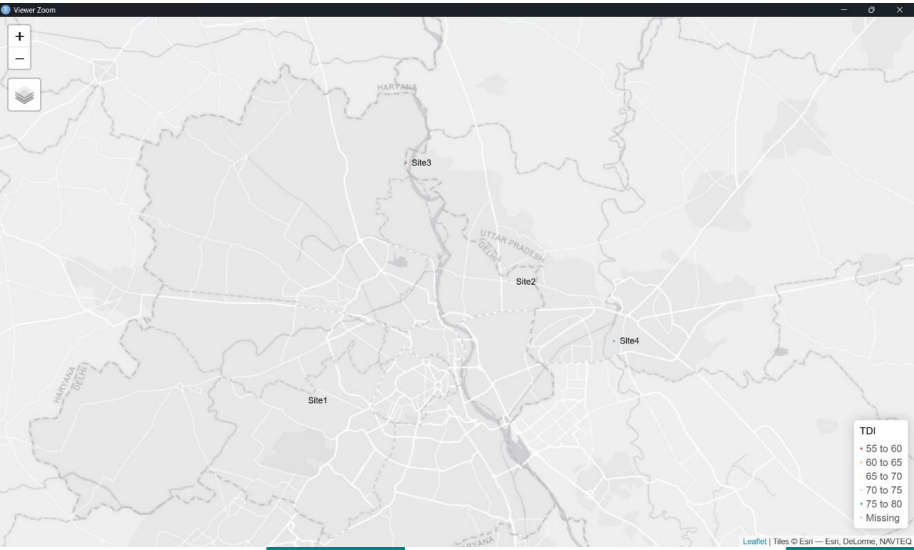


Fig. 10.4 **sf / tmap / leaflet**: Spatial analysis and mapping tools in R, suited for visualizing geographic patterns where maps provide clearer insight than tables.

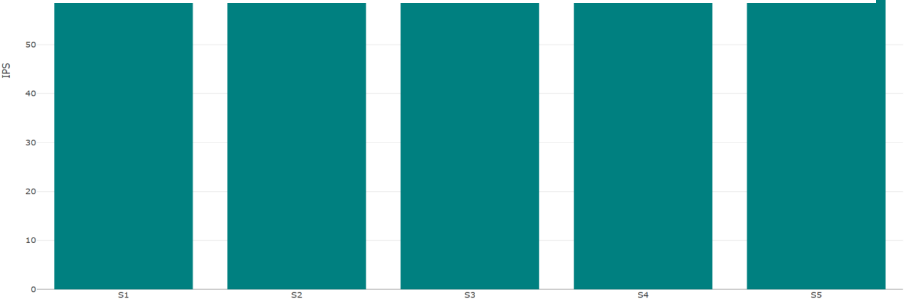


Fig. 10.3 **Plotly**: An R visualization library designed for interactive data exploration, supporting hover information, zooming, and dynamic user interaction.

- **vegan** → for ordination plots, NMDS, PCA, CCA. *#Refer Fig. 10.5*

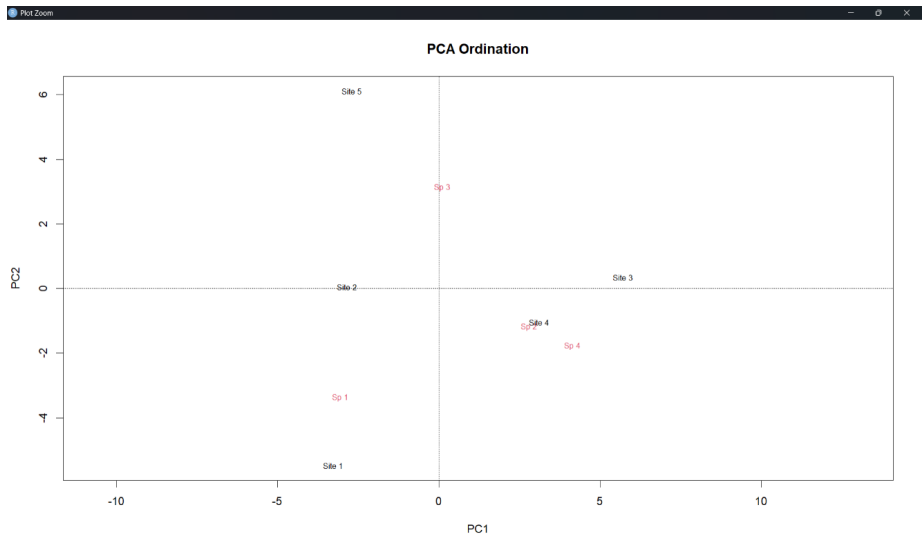


Fig. 10.5 **vegan**: An R package for multivariate ecological analysis, widely used for ordination techniques such as NMDS, PCA, and CCA.

- **radarchart / fmsb** → perfect for multi-index radar plots. *#Refer Fig. 10.6*

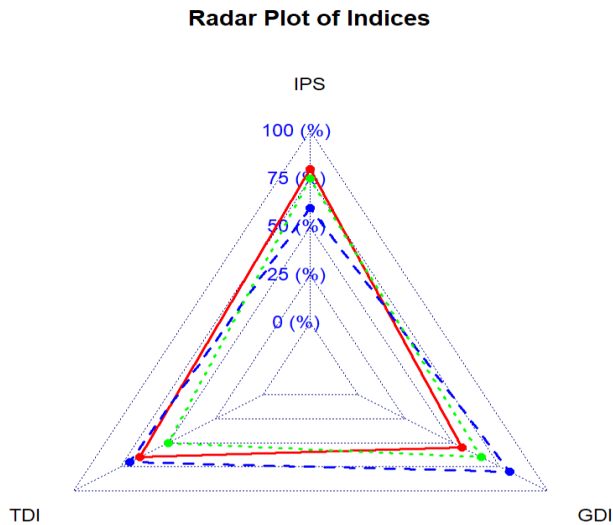


Fig. 10.6 **radarchart / fmsb**: R packages designed for radar chart visualizations, suitable for comparing multiple ecological indices simultaneously.

- **ComplexHeatmap / pheatmap** → when patterns are hidden in giant tables. *#Refer Fig. 10.7*

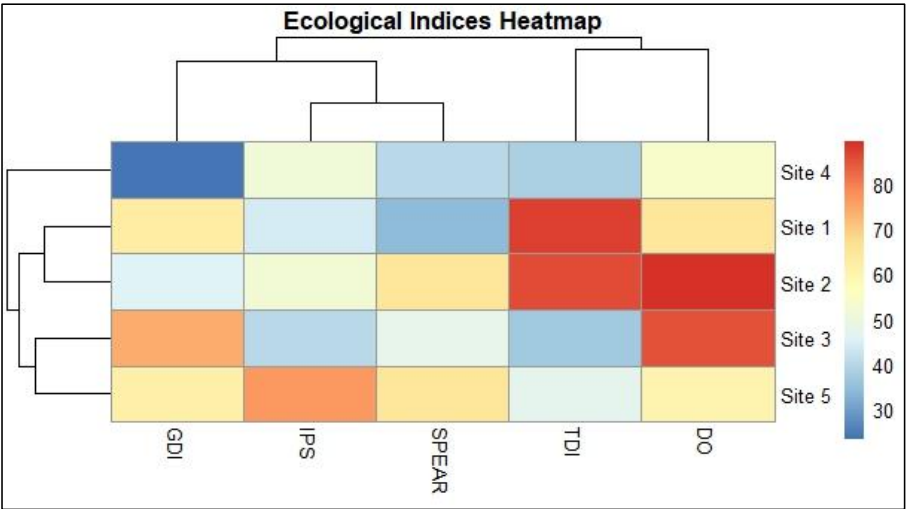


Fig. 10.7 Heatmap generated using `ComplexHeatmap` / `pheatmap`, illustrating how multidimensional data hidden within large tabular datasets can be transformed into clear, interpretable patterns through hierarchical clustering and color gradients.

- `circlize` / `chorddiag` → for relationships, flows, and networks. #Refer Fig. 10.8

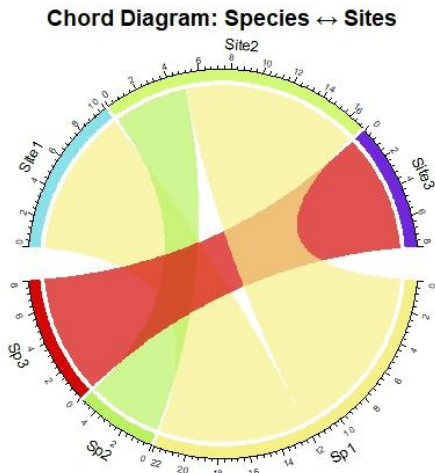


Fig. 10.8 Circular visualization created using `circlize` / `chorddiag`, depicting complex relationships, directional flows, and network interactions among multiple entities through interconnected chords and segments.

Visualization is not decoration. It's translation. It's how numbers leap off the page and into minds. Remember: "The goal is not to impress with complexity, but to illuminate with clarity."

So, the next time you calculate an index in *DiaThor*, don't stop at the table. Ask: how will this look? how will this speak? Because once data is visible, relationships stop hiding. And that's when science becomes insight.

CHAPTER 12

QUICK FIX STEPS FOR COMMONLY ENCOUNTERED ERRORS

"Every error message is not a dead end; it's a signpost pointing to where you've strayed."

ERROR MESSAGE (IN R CONSOLE)	WHAT IT USUALLY MEANS	HOW TO FIX IT
Error in diat_loadData(): object 'Species' not found	The first column isn't named species OR you forgot to include species names.	Rename the first column header to exactly species. Make sure it's the first column.
Some species not recognized	Species names don't match DiaThor's reference database. (e.g., spelling error, space instead of underscore, outdated taxonomy).	Check spelling. Replace spaces with <code>_</code> . Cross-check with DiaThor species list (GitHub/Wiki).
NAs introduced by coercion	Non-numeric values in abundance columns (e.g., present, NA, -, or text).	Replace with numbers (0 if absent). Remove text.
Subscript out of bounds	File has blank rows, hidden columns, or merged cells.	Clean dataset: delete empty rows/columns, ensure each site has one column.
Error: no applicable method for 'mutate'	The package dplyr (or others) wasn't loaded.	Run <code>library(dplyr)</code> before running DiaThor functions.
Error: could not find function "diat_ips"	DiaThor package not loaded OR installed incorrectly.	Run <code>library(DiaThor)</code> . If still fails: <code>devtools::install_github("limnolab/DiaThor")</code> .
Warning: Species not matched	DiaThor couldn't map some taxa to an index.	Check species spelling. Use synonyms if needed.
Error: Input must be a data.frame	Wrong file format (Excel .xlsx or .xls).	Save file as .csv UTF-8.
Error: missing values in abundance data	Blank cells left in site columns.	Replace with 0 or delete the row if not valid.
Graph not appearing	Forgot to save result of function (e.g., <code>results_ips <- diat_ips(df)</code>).	Always assign results to an object (<code>results_*</code>) before plotting.
Plot labels cut off / unreadable	Default R plots don't auto-resize.	Use <code>theme()</code> in <code>ggplot2</code> to increase font size, or adjust width/height in <code>ggsave()</code> .

Error: object 'df' not found	You didn't assign your dataset to df (or you mistyped it).	After loading data, always run: <code>df <- diat_loadData("path")</code> .
Graph looks empty / flat line	Abundance data is all zeros, or index not applicable to your species.	Check dataset with <code>summary(df)</code> . Ensure you have enough species recognized.
Error in View(): object not found	You tried <code>View(results_ips)</code> before saving <code>results_ips</code> .	Always assign: <code>results_ips <- diat_ips(df)</code> then run <code>View(results_ips)</code> .
DiaThor takes too long / freezes	Very large dataset or unnecessary columns.	Keep only Species + sample columns. Remove metadata (dates, GPS, etc.).
Wrong decimal separator	Your CSV uses commas (,) instead of dots (.)	Save CSV with UTF-8 and dot (.) decimals. In Excel: File → Save As → CSV (UTF-8).
Species names too long	Some indices fail when names contain strange characters.	Remove special characters ((), [], ?, etc.) from names.
could not find function "%>%"	The pipe operator from dplyr wasn't loaded.	Run: <code>library(dplyr)</code> .
Colors in plot look bad	Default R colors are not publication-ready.	Add: <code>+ scale_fill_brewer(palette="Set3")</code> or use <code>viridis::scale_color_viridis()</code> .
Labels cut in saved image	Exported image too small.	Save with bigger size: <code>ggsave("plot.png", width=12, height=8, dpi=300)</code> .
Error: package 'DiaThor' not available	R can't find DiaThor on CRAN (it's on GitHub).	Install devtools: <code>install.packages("devtools")</code> then <code>devtools::install_github("limnolab/DiaThor")</code> .
Warnings about NA in results	Some species not recognized in reference tables.	Check species list; accept that some taxa won't be used in index.

Error in <code>diat_loadData()</code> : file not found	Wrong path or slashes in file location.	Use forward slashes / or double backslashes \\. Example: "C:/DIATHOR/data.csv".
CSV won't load	Wrong encoding or hidden spaces in headers.	Save as CSV UTF-8 and remove extra spaces in column names.
Index output gives only zeros	Data not normalized, or all abundances small.	Use raw abundance counts (not percentages) in your matrix.
Some indices missing in output	Your dataset lacks required taxa for those indices.	Not all indices apply to every dataset. Try running a subset.
Error in <code>diat_ips(df)</code>	Input object <code>df</code> not recognized as DiaThor format.	Always use <code>df <- diat_loadData("file.csv")</code> before running indices.
Graph looks squished	Default plot window too small.	Enlarge plot pane in RStudio or use <code>ggsave()</code> with width/height.
Colors all look same	Too many categories.	Add <code>+ scale_fill_viridis_d()</code> or <code>+ scale_color_brewer(palette="Paired")</code> .
Overlapping labels	Too many sites or indices on x-axis.	Rotate labels: <code>+ theme(axis.text.x=element_text(angle=45, hjust=1))</code> .
Error installing DiaThor	GitHub version needs devtools.	Run: <code>install.packages("devtools"); devtools::install_github("limnolab/DiaThor")</code> .
Plots too cluttered	You're plotting all sites/indices at once.	Subset: <code>df_subset <- df[1:10,]</code> and rerun.
Radar plot unreadable	Too many metrics.	Plot only key ones (IPS, TDI, SLA, DDI).

Function not found (diat_xxx)	Typo in function name.	Check with <code>ls("package:Diathor")</code> for exact names.
Script stops halfway	One function failed but others could run.	Use <code>tryCatch()</code> to skip errors and continue analysis.
Missing legend titles	<code>ggplot2</code> hides it if no fill/color.	Add <code>labs(fill="Index Value", color="Index")</code> .
Warning: Removed NAs	Some taxa missing from reference DB.	Acceptable; those species aren't used in calculation.
could not find function "%>%"	Pipe operator not loaded.	Run <code>library(dplyr)</code> or use new R native pipe
Graph not saving	Forgot to use <code>ggsave()</code> .	Example: <code>ggsave("IPS_plot.png", p, dpi=300, width=10, height=7)</code> .
R session crashes	Large dataset overload.	Split into chunks or increase memory. Save intermediate results.
Confused about index meaning	Diathor doesn't explain in plain terms.	Create a lookup table (Index = Name + Range + Interpretation).
Forgotten what's installed	Hard to recall package names.	Run <code>installed.packages()[,1]</code> .
Forgot results variable name	You ran multiple indices.	Use <code>ls(pattern="results_")</code> to list all outputs.
RStudio very slow	Workspace bloated with objects.	Use <code>rm(list=ls())</code> and restart session.
Trouble sharing results	Plots stuck in R.	Export with <code>ggsave()</code> or <code>write.csv(results_ips, "ips.csv")</code> .

Encountering complex error messages in the R console can initially appear discouraging for researchers transitioning from traditional biological methodologies to computational analysis. However, error messages should be viewed as essential diagnostic tools rather than systemic failures. Utilizing open-access software enhances the quality, efficiency, and reproducibility of ecological research, marking a necessary progression from conventional workflows. Computational errors are an inevitable aspect of programming, typically indicating a mismatch between user input and the software's syntax or formatting requirements. The objective is not to eliminate errors entirely, but to systematically decode these outputs and apply appropriate corrections. This requires a methodological shift: treating error outputs as direct, actionable feedback. This chapter outlines systematic troubleshooting steps to effectively resolve common computational issues, ensuring robust and efficient data analysis.

Step 1: Systematically Analyze the Error Output

Rather than immediately resorting to external search engines, researchers should carefully read the console's error output. In most instances, the message directly identifies the root cause, such as a missing parenthesis, an undefined object, or an unloaded package.

Step 2: Verify Syntax and Formatting

Programming languages require strict adherence to syntax. Omitting a comma, bracket, or quotation mark will result in execution failure. Researchers must proofread their scripts methodically to identify typographical errors.

Step 3: Confirm Package and Environment Loading

Execution failures frequently occur when required packages or datasets are not actively loaded into the R environment. It is imperative to verify that all necessary libraries have been initialized before running dependent functions.

Step 4: Execute Code in Modular Segments

Rather than executing an extensive script in its entirety, researchers should run code in smaller, modular sections. This technique isolates variables and makes it significantly easier to pinpoint the exact line causing the computation to fail.

Step 5: Restart the R Session

The R environment can occasionally become cluttered with residual objects from previous analyses, leading to package conflicts or memory exhaustion. Restarting the R session clears the global environment, ensuring a clean workspace for computation.

Step 6: Formulate Precise Troubleshooting Queries

When utilizing external resources or community forums (e.g., GitHub, StackOverflow) for troubleshooting, queries must be highly specific. A well-formulated question should detail the intended objective, the exact code executed, the expected outcome, and the specific error message generated.

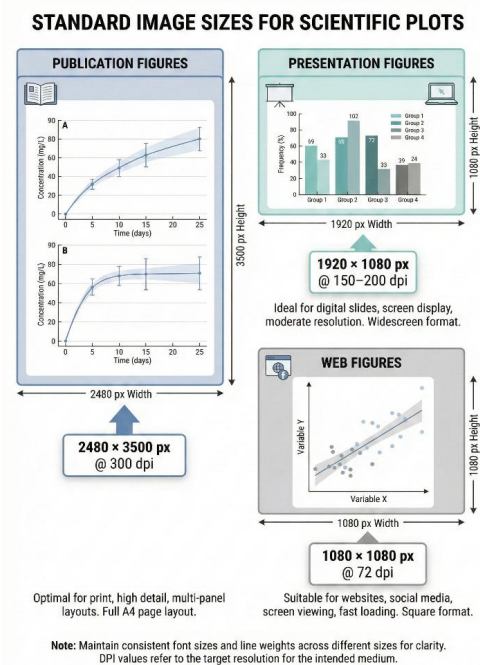
Step 7: Maintain an Analytical Troubleshooting Log

Documenting resolved errors serves as a critical reference for future analyses. Maintaining a systematic log of encountered errors, the context in which they occurred, and their corresponding solutions significantly improves computational fluency and reduces future troubleshooting time. In computational ecology, understanding and resolving programmatic errors is an integral part of ensuring robust, reproducible science.

CHAPTER 13

STANDARD IMAGE SIZE FOR PLOT (IN PIXELS)

"An image is not just a picture; it's a message wrapped in pixels."



Have you ever made a plot with beautiful colors, neat labels, perfect axis ticks and then exported it only to find... it looks blurry? Or worse, the axis text is cut off when you try to paste it into a report or slide. That, my friend, is the moment you realize: pixels matter.

Think of pixels as the bricks of your digital house. The sturdier and more numerous the bricks, the stronger and clearer the picture. If your house is built of too few bricks, it will collapse when stretched across a giant poster. If you overbuild with millions of bricks for a tiny Word figure, your file will get unnecessarily heavy. So, we need a middle ground: standard image sizes that balance clarity with practicality.

Every time you create a plot, you're deciding (knowingly or not) how it will live in the world. Will it appear in a scientific paper? A PowerPoint slide? A website? Each medium speaks a different visual language, and pixels are the vocabulary.

For research papers (print quality): you need higher resolution,

For slides and presentations: clarity matters, but file size is also key (nobody wants a 100 MB PPT that crashes mid-seminar).

For web or social media: small, optimized images rule.

There's no single "perfect" size, but over time, practice has carved out some standard ranges:

Journal figures (high resolution):

300 dpi minimum. That translates to about 2400 × 1800 pixels for a half-page figure. If it's a full-page, you may go 3600 × 2400 pixels.

PowerPoint slides (projector clarity):

1920 × 1080 pixels (HD standard) is usually perfect. It's crisp without being bulky.

Web graphics (fast loading):

Around 1200 × 800 pixels. Enough detail to be sharp, but light enough for quick viewing.

R Plots and Pixel Discipline

In R, you control size using the `png()`, `jpeg()`, or `tiff()` functions.

For example:

Save a plot with controlled pixel size

```
png("plot.png", width = 2400, height = 1800, res = 300)
plot(1:10, 1:10, main = "Pixel Perfect Plot")
dev.off()
```

Here, width and height are in pixels, and res sets the resolution in **dpi (dots per inch)**. If you've ever wondered why your exported plot looked jagged, chances are you left out that `res = 300`.

LABEL	WIDTH (PX)	HEIGHT (PX)	ASPECT RATIO
SMALL (SD)	800	600	4:3
MEDIUM (HD)	1280	720	16:9
LARGE (FULL HD)	1920	1080	16:9
TALL PORTRAIT	1080	1920	9:16
SQUARE	1080	1080	1:1
PRINT- READY	2480	3508	A4 (8.27X11.6 9 IN @300DPI)
POSTER SIZE	3840	2160	16:9 ULTRA HD

Fig. 12.1 Standard plot image size recommended for high-quality visualization and publication, expressed in pixels to ensure clarity, readability, and compatibility across journals, presentations, and digital platforms.

Interactive Moment

Okay, imagine you're making a figure for a journal. You try exporting at 800 × 600 pixels. You zoom in. Letters blur; lines break. The editor sends you a cranky email: "Please resubmit high-resolution images."

Now, you try again at **2400 × 1800 pixels, 300 dpi**.

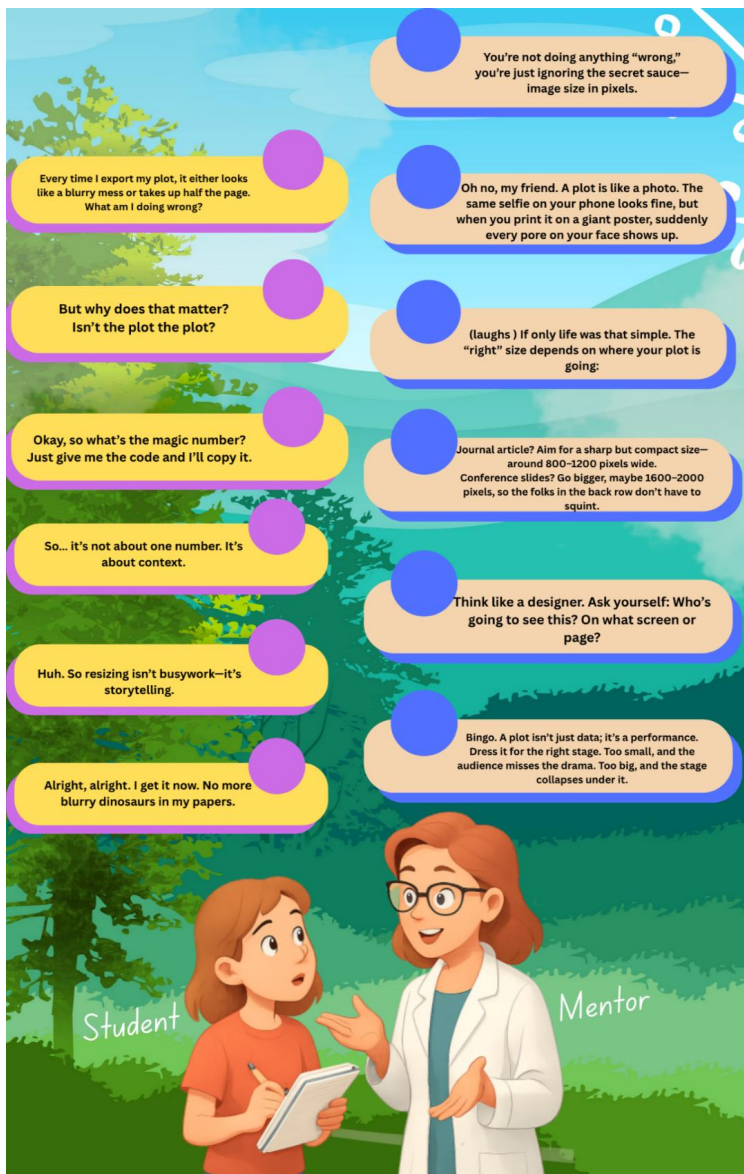
Crisp, professional, print-ready.

See the difference? That's the power of choosing the right pixel size.

Think of this as tailoring a suit. Likewise, you shouldn't use web-sized graphics in a peer-reviewed paper. Each setting demands the right fit.

So, the next time you export a plot, ask yourself: Where is this image going to live? Then pick a pixel size that matches its destiny. *#Refer Fig. 12.1*

Export → Save as Image → Width & Height → Use values from table



CHAPTER 14

DRAWING PLOTS FROM DIATHOR OUTPUTS

“Data may reveal silent numbers, but it begins to speak in a well-crafted graph”

In ecological studies, particularly those involving microscopic organisms like diatoms, raw numerical outputs require visualization to become fully informative. Ecological indices condense biological complexity into numerical values, but plotting these values across multiple samples or time periods is essential for detecting spatial gradients, seasonal trends, and pollutant impacts. A well-designed graph captures relationships that may be overlooked in tabular formats, transforming visualizations into powerful instruments for hypothesis generation and validation. Furthermore, graphical representation enhances the communication of complex ecological data to decision-makers and diverse scientific audiences.

Data Preparation

The DiaThor package outputs its results in tidy data frames, which are optimally structured for visualization in R. In this format, each column represents a discrete variable (e.g., SampleID, Index Value), and each row constitutes a unique observation. This structure is highly compatible with R's powerful visualization libraries, particularly ggplot2 within the tidyverse suite. Prior to generating plots, researchers must verify data integrity by checking for missing values, confirming appropriate data types (such as formatting dates correctly), and ensuring that sample names are meaningful and consistently formatted. Informative titles and axis labels should be utilized to maintain clarity, and abbreviations must be explicitly defined. *#Refer Fig. 13.1*

The act of plotting is not just for communication, it is analytical. Anomalies like extremely low IPS scores or zero guilds can be spotted visually. Clusters in a heatmap might suggest pollution gradients or methodological inconsistencies. When plots are used alongside statistical tools, they become powerful instruments of hypothesis generation and validation. Explaining large datasets to non-specialist audiences or decision-makers can be challenging. Visual representations effectively translate complex diatom indices into accessible formats, such as plots, graphs, and maps using R. That’s where plots come in. In this chapter, we’re going to learn how to let our

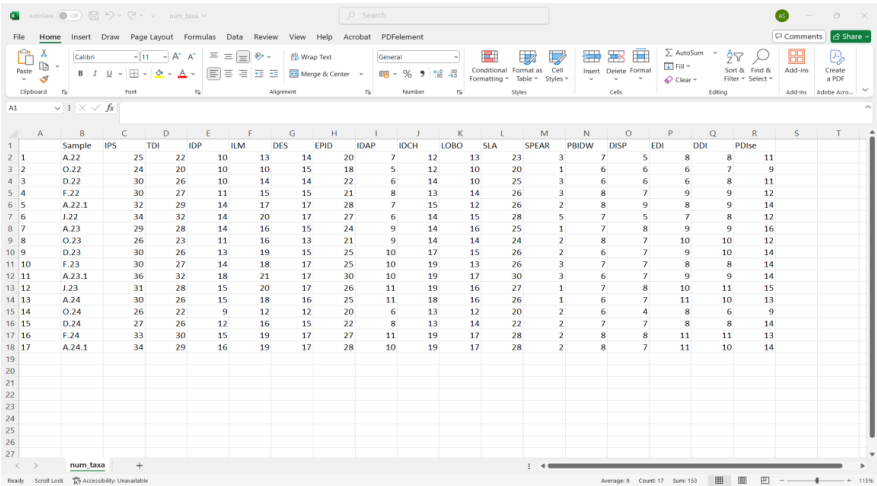


Fig. 13.1 Excel sheets generated from DiaThor outputs, serving as the data source for plotting diatom-based ecological indices and trait distributions used in subsequent visualizations.

diatom indices speak visually, through plots, graphs, and even maps using R.

A standardized workflow utilizing ggplot2 allows for the efficient and reproducible visualization of DiaThor index outputs. The following example demonstrates the primary workflow for generating a bar plot of the Specific Pollution Sensitivity Index (IPS) across multiple sampling sites.

What Does DiaThor Give You?

When you run an index say, `diat_ips()`, it gives you a tidy data frame. As shown in the figure below. That's a fancy way of saying; the table is neat. Each column is a variable. Each row is a sample. Each cell is one value. That's already ready to plot. Now all we need to do is use `ggplot2` and give it instructions like, 'Plot SampleID on x, IPS on y, and draw bars.' Simple! *#Refer Fig. 13.2*

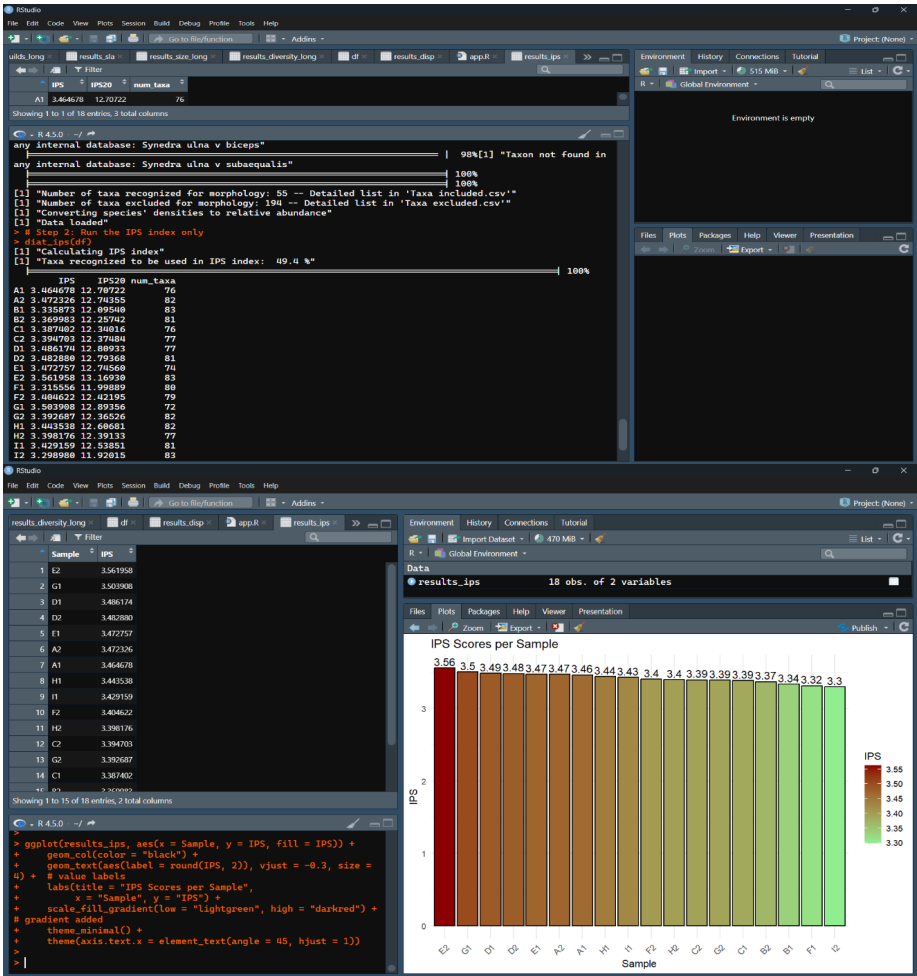


Fig. 13.2 Plots derived from DiaThor outputs, translating diatom-based ecological indices and trait data into clear visual representations that support interpretation of water quality, environmental gradients, and ecosystem health.

PLOTTING WORKFLOW:

#CODE

```

# Load DiaThor
library(diathor)

# Step 1: Load your data (make sure your CSV file is correctly formatted)
df <- diat_loadData(resultsPath = "C:/DIATHOR/CASE 2")

# Step 2: Run the IPS index only
diat_ips(df)

# Step 3: View the results
results_ips <- diat_ips(df)
View(results_ips)

# Step 4: Load required packages
library(ggplot2)
library(dplyr)

# Step 5: Create a sample data frame (use your actual results instead)
results_ips <- data.frame(
  Sample = c("A1", "A2", "B1", "B2", "C1", "C2", "D1", "D2", "E1", "E2",
             "F1", "F2", "G1", "G2", "H1", "H2", "I1", "I2"),
  IPS = c(3.464678, 3.472326, 3.335873, 3.369983, 3.387402, 3.394703,
          3.486174, 3.482880, 3.472757, 3.561958, 3.315556, 3.404622,
          3.503908, 3.392687, 3.443538, 3.398176, 3.429159, 3.298980)
)

# Step 6: Optional: reorder samples by IPS score
results_ips <- results_ips %>%
  arrange(desc(IPS)) %>%
  mutate(Sample = factor(Sample, levels = Sample))

# Step 6: Plot it
ggplot(results_ips, aes(x = Sample, y = IPS, fill = IPS)) +
  geom_col(color = "black") +
  geom_text(aes(label = round(IPS, 2)), vjust = -0.3, size = 4) + # value labels
  labs(title = "IPS Scores per Sample",
       x = "Sample", y = "IPS") +
  scale_fill_gradient(low = "lightgreen", high = "darkred") + # gradient added
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 13.2

```

This gradient bar chart shows:
 Bars with low IPS = light green
 Bars with high IPS = dark red

COLOUR PALLETE OPTIONS

SINGLE GRADIENT OPTIONS

Think of it like a dimmer switch: low light to high light

```
scale_fill_gradient(low = "white",      high = "black")
scale_fill_gradient(low = "lightblue",  high = "darkblue")
scale_fill_gradient(low = "lightgreen", high = "darkgreen")
scale_fill_gradient(low = "yellow",     high = "red")
scale_fill_gradient(low = "pink",       high = "purple")
scale_fill_gradient(low = "beige",      high = "brown")
scale_fill_gradient(low = "ivory",      high = "navy")
scale_fill_gradient(low = "grey90",     high = "grey10")
scale_fill_gradient(low = "cyan",       high = "blue4")
scale_fill_gradient(low = "gold",       high = "firebrick")
```

MULTIPLE GRADIENT OPTIONS

If you've got a midpoint, like a traffic light: green → yellow → red.

```
scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0)
scale_fill_gradient2(low = "darkgreen", mid = "yellow", high = "red", midpoint = 50)
scale_fill_gradient2(low = "navy", mid = "grey90", high = "maroon", midpoint = 10)
scale_fill_gradient2(low = "purple", mid = "pink", high = "orange", midpoint = 5)
scale_fill_gradient2(low = "cyan", mid = "white", high = "brown", midpoint = 20)
scale_fill_gradient2(low = "darkblue", mid = "lightblue", high = "gold", midpoint = 100)
scale_fill_gradient2(low = "black", mid = "grey", high = "white", midpoint = 0)
scale_fill_gradient2(low = "forestgreen", mid = "lightyellow", high = "firebrick",
midpoint = 25)
scale_fill_gradient2(low = "darkred", mid = "white", high = "darkgreen", midpoint = 0)
scale_fill_gradient2(low = "darkorange", mid = "khaki", high = "darkviolet", midpoint =
10)
```

VIRIDIS OPTIONS (CONTINUOUS) OPTIONS

The modern, inclusive palettes. They look classy, work for everyone (even folks with colorblindness), and scale beautifully.

```
scale_fill_viridis_c(option = "viridis") # classic green-blue
scale_fill_viridis_c(option = "magma")   # black-red-yellow
scale_fill_viridis_c(option = "plasma")  # purple-pink-yellow
scale_fill_viridis_c(option = "inferno") # black-orange-yellow
scale_fill_viridis_c(option = "cividis") # yellow-blue (accessible)
scale_fill_viridis_c(option = "turbo")   # rainbowish but smooth
scale_fill_viridis_c(option = "A")       # alt viridis palette A
scale_fill_viridis_c(option = "B")       # alt viridis palette B
scale_fill_viridis_c(option = "C")       # alt viridis palette C
scale_fill_viridis_c(option = "D")       # alt viridis palette D
```

DIVERGING BREWER PALETTES OPTIONS

The old-school heroes, hand-picked by cartographers. If your plot needs to scream "scientific map!", These are the ones.

```
scale_fill_distiller(palette = "RdYlGn", direction = 1)
scale_fill_distiller(palette = "Spectral")
scale_fill_distiller(palette = "PiYG")
scale_fill_distiller(palette = "PRGn")
```

```

scale_fill_distiller(palette = "RdBu")
scale_fill_distiller(palette = "RdGy")
scale_fill_distiller(palette = "PuOr")
scale_fill_distiller(palette = "BrBG")
scale_fill_distiller(palette = "RdYlBu")
scale_fill_distiller(palette = "Set3") # multicolour categorical feel #Refer Fig. 13.3

```

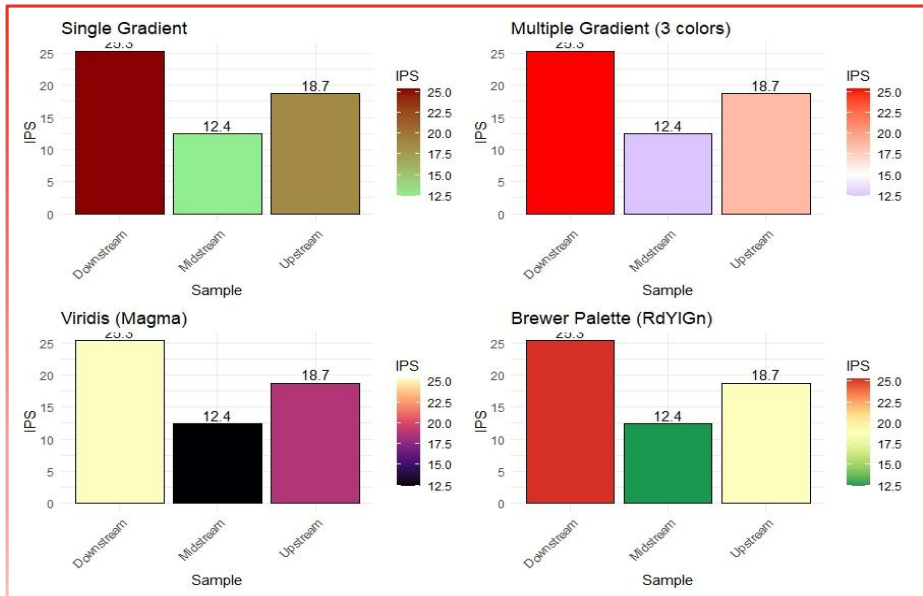


Fig. 13.3 Color palette options illustrating diverse schemes for effective data visualization, enhancing contrast, readability, and ecological interpretation across plots generated from analytical outputs.

To use the full color cheat sheet above, you don't change your whole script ; just swap out that one line depending on what style you want. Everything else stays the same.

```

ggplot(results_ips, aes(x = Sample, y = IPS, fill = IPS)) +
  geom_col(color = "black") +
  geom_text(aes(label = round(IPS, 2)), vjust = -0.3, size = 4) +
  labs(title = "IPS Scores per Sample",
        x = "Sample", y = "IPS") +

```

```

# << Pick ONE of these depending on your gradient choice >>
scale_fill_gradient(low = "lightgreen", high = "darkred") +
  OR
scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 3.4) +
  OR
scale_fill_viridis_c(option = "plasma") +
  OR
scale_fill_distiller(palette = "Spectral") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

This fundamental workflow maps the Sample ID to the x-axis and the computed index score to the y-axis, applying a dynamic color fill based on the index value to emphasize ecological gradients across the sampling sites.

COLOR SELECTION

To maintain visual clarity and academic rigor, color palettes should be selected based on the specific nature of the ecological data being presented. Rather than relying on an extensive list of arbitrary gradients, researchers should utilize a curated set of palettes tailored for scientific interpretation:

Sequential Gradients: Ideal for representing unidirectional variables, such as increasing species richness or progressive nutrient enrichment.

Implementation: `scale_fill_gradient(low = "lightblue", high = "darkblue")`.

Diverging Gradients: Highly effective for indices with established ecological thresholds or midpoints, illustrating transitions from healthy (e.g., blue or green) to degraded (e.g., red or orange) environmental condition

CHAPTER 15

ONE INDEX AT A TIME: CUSTOM ANALYSIS WITH DIATHOR

"Precision isn't found in doing everything at once – it's born from isolating one index and truly understanding what it whispers about the water."

14.1 INTRODUCTION TO MODULAR ANALYSIS

While the `diaThorAll()` function is highly useful for running a comprehensive set of indices and trait-based metrics simultaneously, researchers may occasionally prefer to compute a single index at a time. This targeted approach is particularly beneficial when:

- Troubleshooting or seeking a deeper ecological understanding of a specific index.
- Working with datasets tailored strictly for specific analytical metrics.
- Performing comparative analyses across multiple runs with altered parameters.

DiaThor is modular by design. Each index within the package has a dedicated function that can be executed individually, yielding focused outputs specific to that particular index or ecological trait group.

14.2 DATA PREPARATION FOR INDIVIDUAL INDICES

To execute a single index analysis, users must adhere to a sequential workflow:

Load the sample data: Ensure the CSV file is accurately formatted.

Run the `diat_loadData()` function: This standardizes and prepares the dataset.

Execute the desired index function: Apply the specific command (e.g., `diat_ips()`) to the processed data object.

Each function accepts a `diat_loadData()` object as input and outputs a tidy data frame containing the computed index scores alongside relevant ecological details.

Best Practices for Individual Execution:

Always validate that species names are taxonomically correct and formatted precisely to avoid execution failures. If an index fails to compute, verify whether the input species are present in that specific index's internal reference database.

Utilize the `function_name` command in R (e.g., `diat_ips`) to access official documentation and syntax examples. Running a single index in DiaThor provides analytical flexibility, allowing researchers to inspect ecological conditions selectively. It proves especially useful during methodological development, error resolution, or when focusing on an environmental stressor highly relevant to a specific region.

#Install DiaThor (if not already done) using

```
install.packages("diathor")
```

Load DiaThor

```
library(diathor)
```

Step 1: Load your data (make sure your CSV file is correctly formatted)

```
df <- diat_loadData(resultsPath = "C:/DIATHOR/CASE 2")
```

Step 2: Run the IPS index only

```
diat_ips(df)
```

Step 3: View the results

```
results_ips <- diat_ips(df)
```

```
View(results_ips)
```

NOTE : FOLLOW THE FORMAT FOR CALCULATING DIFFERENT INDEX

```
diat_<index>(df)
```

```
results_<index> <- diat_<index>(df)
```

```
View(results<index>)
```

#EXPLANATION

```
df <-
```

You're assigning the output to a variable named df.

#This will hold the processed data that DiaThor can use to calculate indices (like IPS, DDI, TDI, etc.)

#Think of df as your ready-to-go diatom abundance table, now formatted and read by DiaThor

```
diat_loadData(...)
```

This is a DiaThor helper function that reads and prepares your data.

It loads either: your raw abundance data,

OR

the output of a previous DiaThor analysis, depending on what's in the path you give it.

Under the hood, it: Looks for files in the specified folder; Imports and validates your data (checks for the correct structure: species as rows, samples as columns, etc.); Converts it into a DiaThor-readable object (usually a list or data frame, depending on use).

```
resultsPath = "C:/DIATHOR/CASE 2"
```

This tells DiaThor where to look; "C:/DIATHOR/CASE 2" is your file path on your computer.

It should contain your input data file (usually a .csv file), named something like: data.csv or results.csv if it's already been processed once.

If DiaThor throws an error like "no valid file found" — check spelling, path format, or file name.

Say you have a dataset exported from Excel named CASE2.csv inside the folder C:/DIATHOR/CASE 2/.

```
df <- diat_loadData(resultsPath = "C:/DIATHOR/CASE 2")
```

will find that file, import it, clean/structure it internally

```
diat_ips(df)
```

```
diat_ddi(df)
```

(Common Mistakes):

Missing .csv file	File not found error
Wrong format inside CSV	Input not valid
Extra headers or blank rows	Structure issues
File open in Excel during import	Locked file error

Step 4: Load required packages

```
library(ggplot2)
library(dplyr)
```

Step 5: Create a sample data frame (use your actual results instead)

```
results_ips <- data.frame(
  Sample = c("A1", "A2", "B1", "B2", "C1", "C2", "D1", "D2", "E1", "E2",
    "F1", "F2", "G1", "G2", "H1", "H2", "I1", "I2"),
  IPS = c(3.464678, 3.472326, 3.335873, 3.369983, 3.387402, 3.394703,
    3.486174, 3.482880, 3.472757, 3.561958, 3.315556, 3.404622,
    3.503908, 3.392687, 3.443538, 3.398176, 3.429159, 3.298980)
)
```

Step 6: Optional: reorder samples by IPS score

```
results_ips <- results_ips %>%
  arrange(desc(IPS)) %>%
  mutate(Sample = factor(Sample, levels = Sample))
```

Step 7: Plot it

14.3 VISUALISING IPS INDEX: GENERATING A BAR GRAPH

```
ggplot(results_ips, aes(x = Sample, y = IPS, fill = IPS)) +
  geom_col(color = "black") +
  scale_fill_gradient(low = "#fef0d9", high = "#a63603") + # warm gradient
  geom_hline(yintercept = 3.5, linetype = "dashed", color = "darkgreen") +
  labs(title = "Diatom-based IPS Scores by Sample",
       subtitle = "Water quality based on diatom pollution sensitivity index",
       x = "Sample ID", y = "IPS Score") +
  theme_minimal(base_size = 14) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none") #Refer Fig. 14.1
```

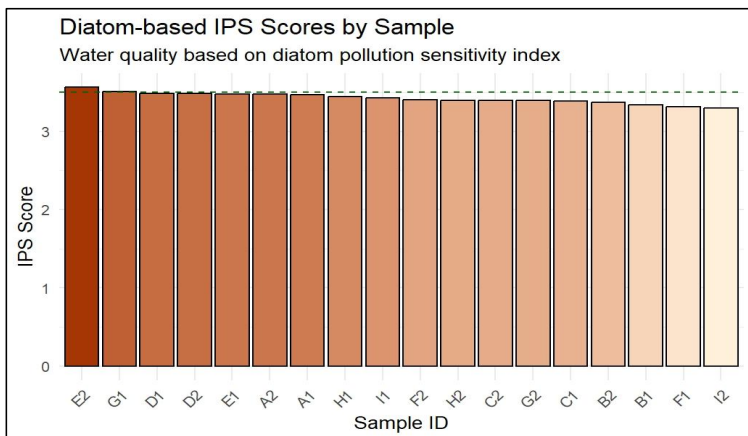


Fig. 14.1 Bar graph illustrating the IPS (Specific Pollution Sensitivity Index), generated from DiaThor outputs to compare site-wise variations in water quality and diatom-based ecological status.

#EXPLANATION

```
library(ggplot2)
library(dplyr)
```

Loads the libraries you'll need:

ggplot2: for plotting

dplyr: for data wrangling

Create a sample data frame (use your actual results instead)

```
results_ips <- data.frame(
  Sample = c("A1", "A2", "B1", "B2", "C1", "C2", "D1", "D2", "E1", "E2",
             "F1", "F2", "G1", "G2", "H1", "H2", "I1", "I2"),
  IPS = c(3.464678, 3.472326, 3.335873, 3.369983, 3.387402, 3.394703,
          3.486174, 3.482880, 3.472757, 3.561958, 3.315556, 3.404622,
          3.503908, 3.392687, 3.443538, 3.398176, 3.429159, 3.298980)
)
```

Creates a data frame named **results_ips** manually. This mimics your DiaThor output.

Sample = your sample labels (A1 to I2).

IPS = IPS values per sample.

```
results_ips <- results_ips %>%
```



```

    arrange(desc(IPS)) %>%
    mutate(Sample = factor(Sample, levels = Sample))
  arrange(desc(IPS)): sorts samples from highest IPS (best water quality) to lowest.
  mutate(Sample = factor(...)): reorders the Sample factor levels so the bar plot doesn't jumble the bars. Bar
    plots respect factor levels.
  Without this, bars would plot in alphabetical order (A1, A2, B1...), not in ranking order.
  ggplot(results_ips, aes(x = Sample, y = IPS, fill = IPS)) +
    Initializes the ggplot. Tells it:
    x-axis = Sample (categorical).
    y-axis = IPS score.
  fill = IPS means bar color will be based on IPS value (i.e., dynamic shading).
    geom_col(color = "black") +
      Draws bars (columns) for each sample; Outlined in black for contrast
  scale_fill_gradient(low = "#f0d9", high = "#a6603") + # warm gradient
    Sets a color gradient for the bars:
    Low IPS = light peach
    High IPS = dark rust-orange
  Gives a beautiful 'heatmap' feel across your bars.* You can switch to cool colors (blues/greens) if you're tired of
    looking at rust.
  geom_hline(yintercept = 3.5, linetype = "dashed", color = "darkgreen") +
    Draws a horizontal reference line at IPS = 3.5, dashed and green.
    Acts as a visual threshold for "pretty good water quality".
    Makes interpretation much faster in presentations.
  labs(title = "Diatom-based IPS Scores by Sample",
    subtitle = "Water quality based on diatom pollution sensitivity index",
    x = "Sample ID", y = "IPS Score") +
    Sets titles and axis labels:
    title: Main plot title
    subtitle: Adds context
    x and y: self-explanatory
  theme_minimal(base_size = 14) +
    Applies a clean and modern plot theme, with a slightly larger font for readability (esp. in publications or slides).
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none")
    Rotates the x-axis labels so they don't overlap.
  legend.position = "none" hides the color legend (because the fill is redundant—you're already showing the value
    directly).

```

LET'S DO SOME CREATIVITY WITH THE BAR GRAPH

#CODE

```

library(ggplot2)

# Assuming your data frame is called results_ips and has 'Sample' and 'IPS' (with visible values above the bar)
ggplot(results_ips, aes(x = Sample, y = IPS)) +
  geom_col(fill = "yellow", color = "black") +
  geom_text(aes(label = round(IPS, 2)), vjust = -0.3, size = 4) + # <- value labels here
  labs(title = "IPS Scores per Sample",
    x = "Sample", y = "IPS") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.2

```

#EXPLANATION

```

aes(label = round(IPS, 2)) → formats IPS values to 2 decimal places.
vjust = -0.3 → shifts labels slightly above the bars.
size = 4 → controls text size (you can tweak it if needed).

```

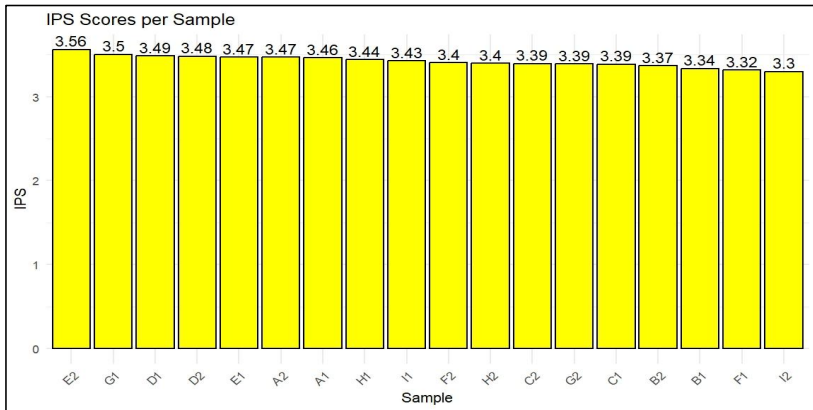


Fig. 13.2 Colored bar graph of the IPS (Specific Pollution Sensitivity Index), using gradient-based color encoding

NOTE: For bold or colored labels:

```
geom_text(aes(label = round(IPS, 2)), vjust = -0.3, size = 4, fontface = "bold",
color = "darkblue")
```

Correct (Modify the complete code as shown:)

```
ggplot(results_ips, aes(x = Sample, y = IPS)) +
```

```
geom_col(fill = "pink", color = "orange") +
```

```
geom_text(aes(label = round(IPS, 2)), vjust = -0.3, size = 4, fontface = "bold",
color = "darkblue") + # <- value labels here
```

```
labs(title = "IPS Scores per Sample",
```

```
x = "Sample", y = "IPS") +
```

```
theme_minimal() +
```

```
theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.3
```

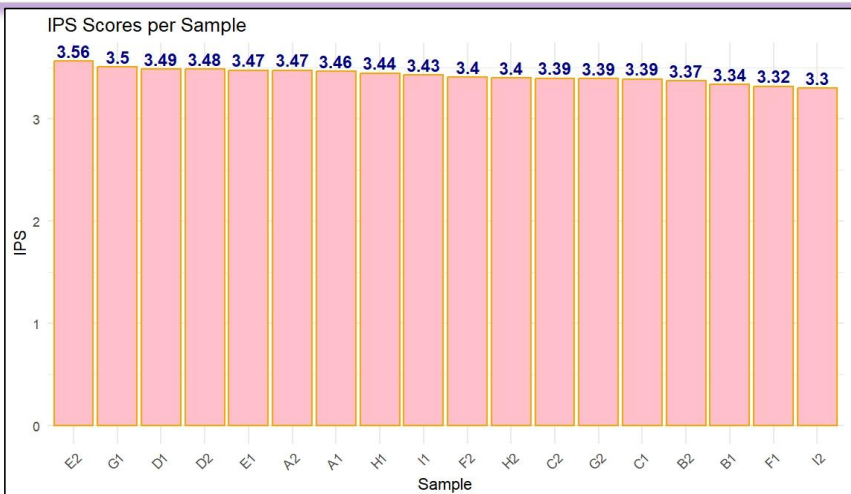


Fig. 14.3 Colored bar graph of the IPS (Specific Pollution Sensitivity Index), using gradient-based color encoding.

14.4 VISUALISING IPS INDEX: GENERATING A LINE GRAPH

```
geom_text(aes(label = round(IPS, 2)), vjust = -1, size = 3.5)
library(ggplot2)

ggplot(results_ips, aes(x = Sample, y = IPS, group = 1)) +
  geom_line(color = "steelblue") +
  geom_point(size = 4, color = "black") +
  geom_text(aes(label = round(IPS, 2)), vjust = -1, size = 3.5) +
  labs(title = "IPS Scores per Sample (Line Plot)",
       x = "Sample", y = "IPS") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.4
```

#EXPLANATION

`geom_text(...)` → places the value labels.

`label = round(IPS, 2)` → shows only 2 decimal places.

`vjust = -1` → moves the text slightly above the dots. If you will write +1 here the values will be shown below the dots

`size = 3.5` → sets readable label size.

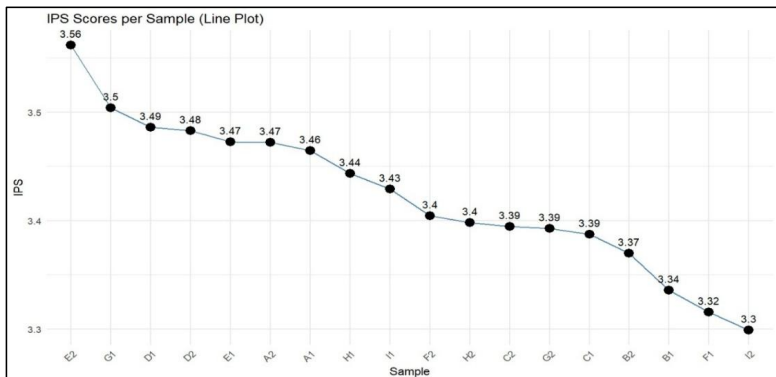


Fig. 14.4 Line graph representing the IPS (Specific Pollution Sensitivity Index) across sampling sites, highlighting temporal or spatial trends in diatom-based water quality assessment

NOTE: For bold or colored labels:

```
geom_text(aes(label = round(IPS, 2)), vjust = -1.3, size = 4, fontface = "bold", color = "red")
```

Correct (Modify the complete code as shown:)

```
ggplot(results_ips, aes(x = Sample, y = IPS, group = 1)) +
```

```
  geom_line(color = "pink") +
```

```
  geom_point(size = 4, color = "brown") +
```

```
  geom_text(aes(label = round(IPS, 2)), vjust = -1.3, size = 4, fontface = "bold", color = "red") +
```

```
  labs(title = "IPS Scores per Sample (Line Plot)",
```

```
       x = "Sample", y = "IPS") + theme_minimal() +
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.5
```

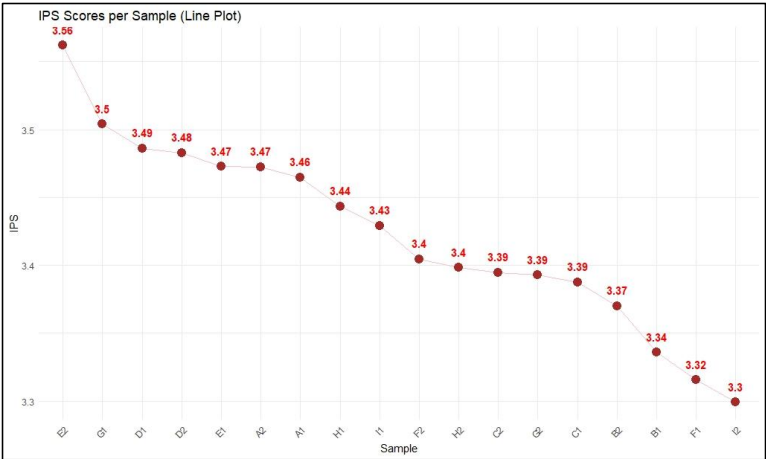


Fig. 14.5 Line graph representing the IPS (Specific Pollution Sensitivity Index) across sampling sites, highlighting temporal or spatial trends in diatom-based water quality assessment (with color changes)

14.5 ADVANCED CUSTOMIZATION: VISUALIZING VAN DAM TRAITS

#Step 1: Tidy Your Data (Long Format)

We'll first convert the wide format to long for plotting:

You can see the result generated by downloading the excel file from the link

<https://github.com/Ar108-bot/DiaThor-Results/commit/2a3a51a6cc30b61427be2c7f56a2de3d4a2e4501#diff-e4cf7ad1676fd4cd78f7bb634417c38329afb0024fe95d8bfde4e991d5d3ca20>

```
library(tidyverse)
```

```
# Assuming your Van Dam result is saved in results_vandam
vandam_salinity <- results_vandam %>%
  select(starts_with("VD.Salinity.")) %>%
  mutate(Sample = paste0("S", row_number())) %>% # Create sample IDs
  pivot_longer(cols = starts_with("VD.Salinity."),
               names_to = "Salinity_Class",
               values_to = "Percentage")
```

Step 2: Plot a Stacked Bar Plot (Salinity Classes per Sample)

```
ggplot(vandam_salinity, aes(x = Sample, y = Percentage, fill = Salinity_Class)) +
  geom_bar(stat = "identity") +
  labs(title = "Van Dam Salinity Classes per Sample",
       x = "Sample", y = "Percentage (%)") +
  scale_fill_brewer(palette = "YlGnBu") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.6
```

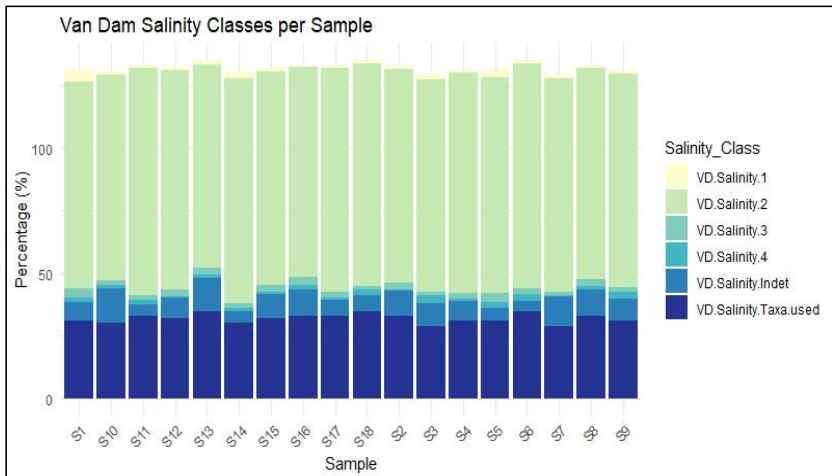


Fig. 14.6 Stacked bar plot showing the distribution of salinity preference classes across samples, illustrating the relative contribution of each class per site.

```
library(tidyverse)
```

```
#Step 1: Use your actual dataframe from diat_vandam(df)
```

```
results_vandam <- diat_vandam(df) # Already run by you
```

```
#Step 2: Add Sample column
```

```
results_vandam <- results_vandam %>%
  mutate(Sample = paste0("S", row_number()))
```

SALINITY CLASS PLOT

```
vandam_salinity <- results_vandam %>%
  select(Sample, starts_with("VD.Salinity."), -VD.Salinity.Taxa.used) %>%
  pivot_longer(cols = -Sample, names_to = "Salinity_Class", values_to = "Percentage")
```

```
ggplot(vandam_salinity, aes(x = Sample, y = Percentage, fill = Salinity_Class)) +
  geom_bar(stat = "identity") +
  labs(title = "Van Dam Salinity Classes", x = "Sample", y = "Percentage (%)") +
  scale_fill_brewer(palette = "YlGnBu") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.7
```

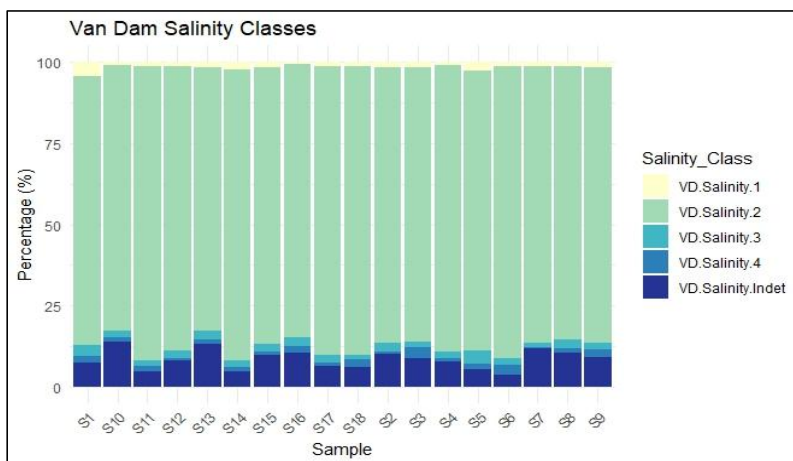


Fig. 14.7 Stacked bar plot showing the distribution of salinity preference classes across samples

OXYGEN REQUIREMENT PLOT

```
vandam_oxygen <- results_vandam %>%
  select(Sample, starts_with("VD.Oxygen."), -VD.Oxygen.Taxa.used) %>%
  pivot_longer(cols = -Sample, names_to = "Oxygen_Class", values_to = "Percentage")
```

```
ggplot(vandam_oxygen, aes(x = Sample, y = Percentage, fill = Oxygen_Class)) +
  geom_bar(stat = "identity") +
  labs(title = "Van Dam Oxygen Requirements", x = "Sample", y = "Percentage (%)") +
  scale_fill_brewer(palette = "Blues") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.8
```




Fig. 14.8 Oxygen requirement plot illustrating the distribution of diatom taxa across different oxygen preference classes for each sample, highlighting variations in dissolved oxygen conditions

SAPROBITY CLASS PLOT

```
vandam_saprobity <- results_vandam %>%
  select(Sample, starts_with("VD.Saprobity."), -VD.Saprobity.Taxa.used) %>%
  pivot_longer(cols = -Sample, names_to = "Saprobity_Class", values_to = "Percentage")

ggplot(vandam_saprobity, aes(x = Sample, y = Percentage, fill = Saprobity_Class)) +
  geom_bar(stat = "identity") +
  labs(title = "Van Dam Saprobity Classes", x = "Sample", y = "Percentage (%)") +
  scale_fill_brewer(palette = "Oranges") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.9
```

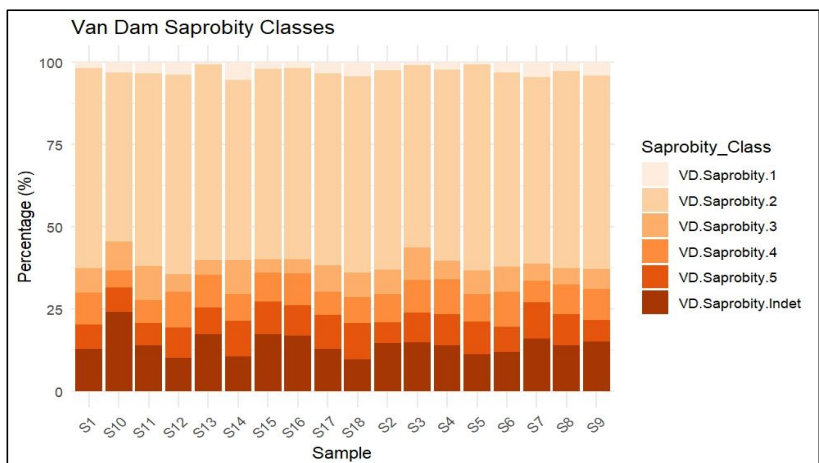


Fig. 14.9 Saprobity class plot depicting the relative abundance of diatom taxa across saprobic categories for each sample, reflecting organic pollution levels and ecological status.

TROPHIC STATE PLOT

```
vandam_trophic <- results_vandam %>%
  select(Sample, starts_with("VD.Trophic."), -VD.Trophic.Taxa.used) %>%
  pivot_longer(cols = -Sample, names_to = "Trophic_Class", values_to = "Percentage")
```

```
ggplot(vandam_trophic, aes(x = Sample, y = Percentage, fill = Trophic_Class)) +
  geom_bar(stat = "identity") +
  labs(title = "Van Dam Trophic State Classes", x = "Sample", y = "Percentage (%)") +
  scale_fill_brewer(palette = "Greens") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.10
```

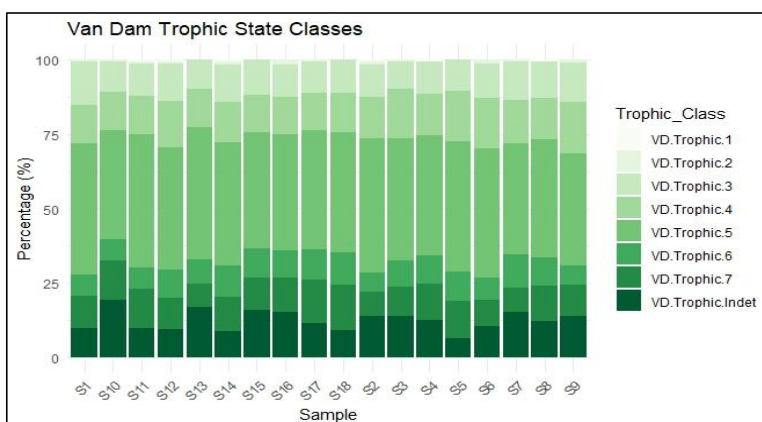


Fig. 14.10 Trophic state plot showing the distribution of diatom taxa across trophic preference classes for each sample, indicating nutrient enrichment levels and ecosystem productivity.

N-HETEROTROPHY PLOT

```
vandam_nhet <- results_vandam %>%
  select(Sample, starts_with("VD.N.Het."), -VD.N.Het.Taxa.used) %>%
  pivot_longer(cols = -Sample, names_to = "N_Het_Class", values_to = "Percentage")

ggplot(vandam_nhet, aes(x = Sample, y = Percentage, fill = N_Het_Class)) +
  geom_bar(stat = "identity") +
  labs(title = "Van Dam N-Heterotrophy Classes", x = "Sample", y = "Percentage (%)") +
  scale_fill_brewer(palette = "Purples") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.11
```

AEROPHILY PLOT

```
vandam_aero <- results_vandam %>%
  select(Sample, starts_with("VD.Aero."), -VD.Aero.Taxa.used) %>%
  pivot_longer(cols = -Sample, names_to = "Aero_Class", values_to = "Percentage")

ggplot(vandam_aero, aes(x = Sample, y = Percentage, fill = Aero_Class)) +
  geom_bar(stat = "identity") +
  labs(title = "Van Dam Aerophily Classes", x = "Sample", y = "Percentage (%)") +
  scale_fill_brewer(palette = "Reds") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) #Refer Fig. 14.12
```

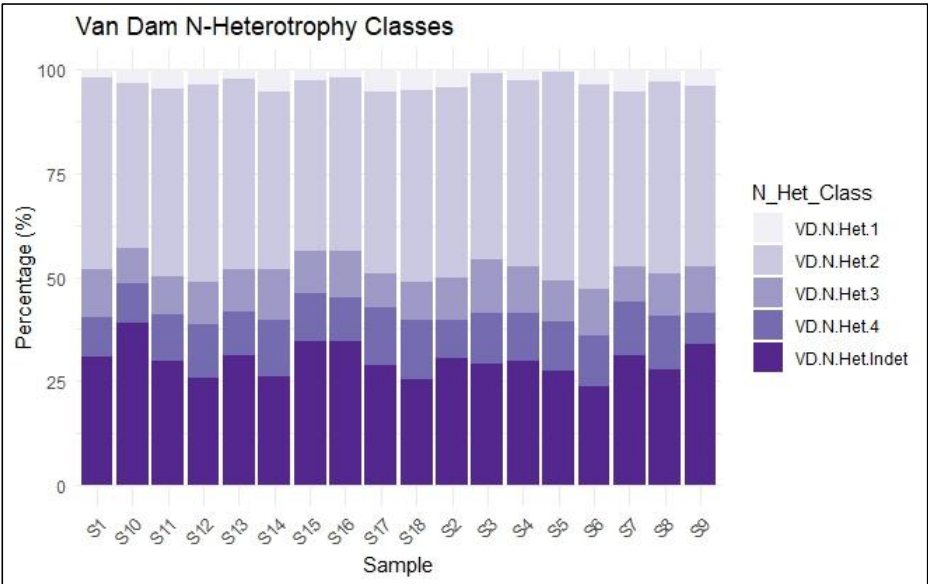


Fig. 14.11 N-heterotrophy plot illustrating the distribution of diatom taxa across nitrogen heterotrophy classes for each sample, reflecting nutrient utilization strategies and nitrogen availability.

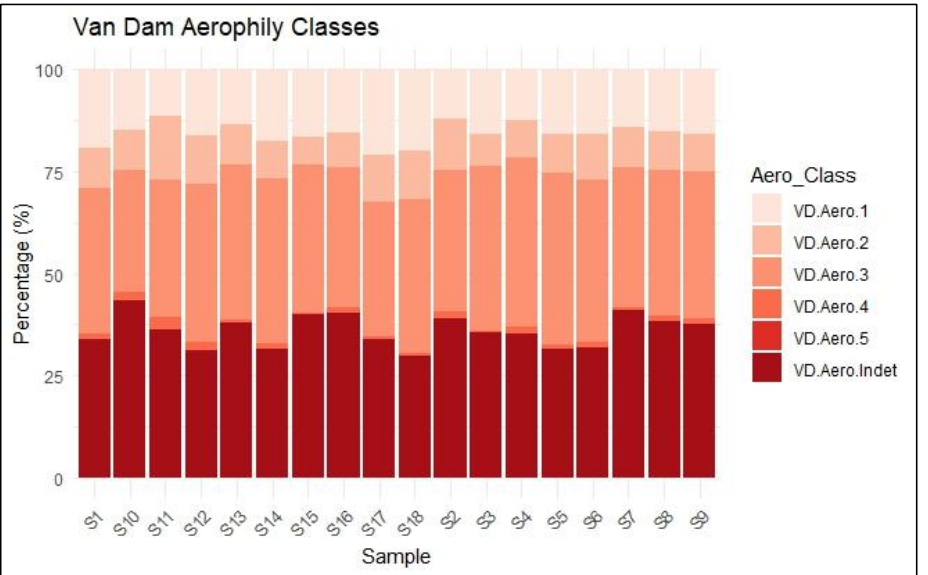


Fig. 14.12 Aerophily plot showing the distribution of diatom taxa across aerophilic preference classes for each sample, indicating tolerance to aerial exposure and hydrological fluctuation.

14.6 AUTOMATING VISUALIZATIONS VIA REUSABLE FUNCTIONS MASTER TEMPLATE (FOR EACH VANDAM TRAIT)

Let's start with the **template code** we used:

```
library(tidyverse) # Load all necessary packages

# Prepare the data

trait_data <- results_vandam %>%
  select(Sample, starts_with("VD.Salinity")) %>%
  pivot_longer(-Sample, names_to = "Class", values_to = "Percentage")

# Clean the class names

trait_data$Class <- gsub("VD.Salinity.", "", trait_data$Class)

# Create the plot

ggplot(trait_data, aes(x = Sample, y = Percentage, fill = Class)) +
  geom_bar(stat = "identity") + # 4a: Actual stacked bar
  theme_minimal(base_size = 14) + # 4b: Clean theme
  theme(axis.text.x = element_text(angle = 45, hjust = 1), # 4c: Tilt x-axis labels
        axis.title = element_text(face = "bold")) + # 4d: Bold axis titles
  labs(title = "Salinity Class Distribution",
       x = "Sample", y = "Percentage (%)")
```

Trait	Replace with...
Oxygen Requirement	"VD.Oxygen"
Saprobity	"VD.Saprobity"
Trophic State	"VD.Trophic"
Heterotrophy	"VD.N.Het"
Aerophily	"VD.Aero"

EXAMPLE FOR TROPHIC STATE:

```
trait_data <- results_vandam %>%
  select(Sample, starts_with("VD.Trophic")) %>%
  pivot_longer(-Sample, names_to = "Class", values_to = "Percentage")

trait_data$Class <- gsub("VD.Trophic.", "", trait_data$Class)

ggplot(trait_data, aes(x = Sample, y = Percentage, fill = Class)) +
  geom_bar(stat = "identity") +
  theme_minimal(base_size = 14) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title = element_text(face = "bold")) +
  labs(title = "Trophic State Classes", x = "Sample", y = "Percentage (%)")
```

#EXPLANATION

Loads a set of packages like dplyr, ggplot2, and tidyr that help us clean and visualize data.

```
trait_data <- results_vandam %>% ...
```

This whole block does three things:

a) results_vandam %>%

Pipe in the dataframe you already created from `diat_vandam(df)`.

```
b) select(Sample, starts_with("VD.Salinity"))
```

Picks only the columns starting with VD.Salinity, plus the Sample column.
You can replace "VD.Salinity" with "VD.Oxygen", "VD.Trophic", "VD.Aero", etc.

```
pivot_longer(-Sample, names_to = "Class", values_to = "Percentage")
```

Transforms wide data (multiple columns for classes) into long format.
You now get two columns:

- ✓ Class – the original column name (like 'VD.Salinity.1')
- ✓ Percentage – the value (like 45.3)

```
trait_data$Class <- gsub("VD.Salinity.", "", trait_data$Class)
```

Cleans the class names by removing the prefix "VD.Salinity."
So "VD.Salinity.1" becomes "1", "VD.Salinity.Indet" becomes "Indet", etc.

This keeps your plot labels neat and short.

Plotting the Stacked Bar

```
ggplot(trait_data, aes(x = Sample, y = Percentage, fill = Class))
```

`x = Sample` → each bar = one sample
`y = Percentage` → height of bar = total percentage of trait categories
`fill = Class` → colors split by each trait class (e.g., Salinity 1 to 4, Indet)

```
4a) geom_bar(stat = "identity")
```

Actually, draws the stacked bars using your data's real values.
`stat = "identity"` tells ggplot to use actual Percentage values instead of counting.

```
4b) theme_minimal(base_size = 14)
```

Applies a clean, modern theme with readable font size (14).

```
4c) theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Rotates x-axis labels by 45° to prevent overlapping text.
Makes things legible when you have many samples.

```
4d) axis.title = element_text(face = "bold")
```

Makes axis titles bold for better visibility.

```
4e) labs(...)
```

Sets your title, x-axis label, and y-axis label.



Fig.14.13 Example of a trophic state plot derived from DiaThor outputs, demonstrating the distribution of diatom taxa across oligotrophic to eutrophic preference classes for a representative sample.

14.7 ASSIGNING DESCRIPTIVE ECOLOGICAL LABELS

```
library(tidyverse)
```

Step 1: Pivot Trophic data

```
trophic_data <- results_vandam %>%
  select(Sample, starts_with("VD.Trophic.")) %>%
  pivot_longer(-Sample, names_to = "Class", values_to = "Percentage")
```

Step 2: Remove "Taxa.used"

```
trophic_data <- trophic_data %>%
  filter(!Class %in% c("VD.Trophic.Taxa.used"))
```

Step 3: Simplify Class Names

```
trophic_data$Class <- gsub("VD.Trophic.", "", trophic_data$Class)
```

Step 4: Add labels (map class numbers to ecological meaning)

```
trophic_labels <- c(
  "1" = "Oligotraphentic",
  "2" = "Oligo-mesotraphentic",
  "3" = "Mesotraphentic",
  "4" = "Meso-eutraphentic",
  "5" = "Eutraphentic",
  "6" = "Hyper-eutraphentic",
  "7" = "Polysaprobic",
  "Indet" = "Indeterminate"
)
```

Step 5: Plot

```
ggplot(trophic_data, aes(x = Sample, y = Percentage, fill = Class)) +
  geom_bar(stat = "identity", color = "white") +
  scale_fill_manual(
    values = c(
      "1" = "#1b9e77", "2" = "#d95f02", "3" = "#7570b3", "4" = "#e7298a",
      "5" = "#66a61e", "6" = "#e6ab02", "7" = "#a6761d", "Indet" = "grey50"
    ),
    labels = trophic_labels,
    name = "Trophic Class"
  ) +
  labs(
    title = "Trophic State Distribution of Diatom Communities",
    subtitle = "Based on Van Dam Classification Scheme",
    x = "Sample", y = "Percentage (%)"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title = element_text(face = "bold"),
    legend.title = element_text(face = "bold")
  )
```

Interpretation

CLASS (LABEL)	ECOLOGICAL MEANING
OLIGOTRAPHENTIC (1)	<i>Very low nutrient waters</i>
OLIGO-MESOTRAPHENTIC (2)	<i>Low to moderate nutrient conditions</i>
MESOTRAPHENTIC (3)	<i>Moderate nutrient levels</i>
MESO-EUTRAPHENTIC (4)	<i>Moderate to high eutrophication</i>
EUTRAPHENTIC (5)	<i>High nutrient availability</i>
HYPER-EUTRAPHENTIC (6)	<i>Extremely nutrient-rich, often polluted</i>
POLYSAPROBIC (7)	<i>Heavily polluted, organic-rich waters</i>
INDETERMINATE	<i>Species with unknown trophic preference</i>

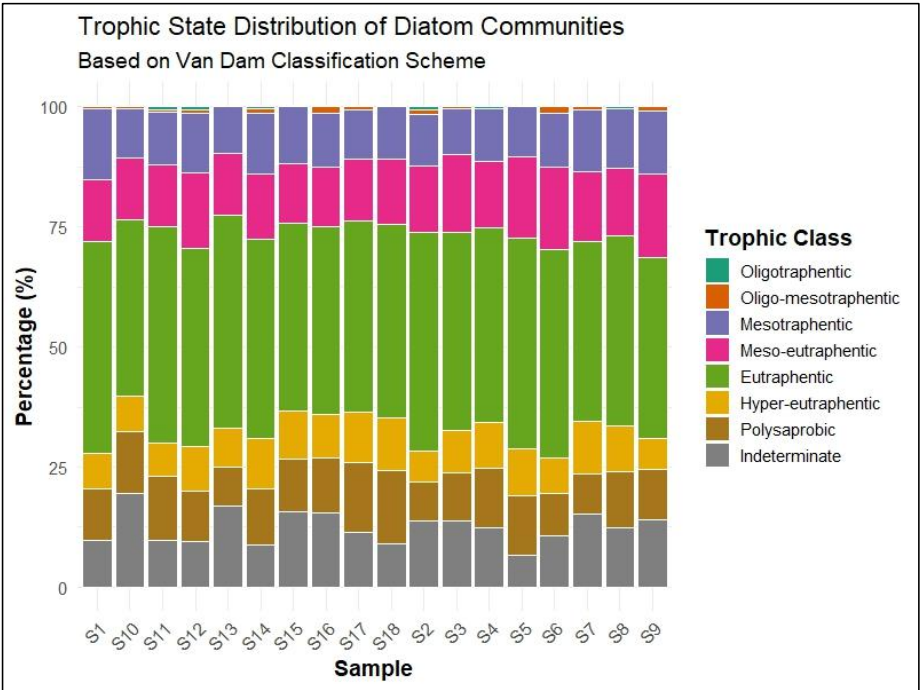


Fig. 14.14 Example of a trophic state plot derived from *DiaThor* outputs, demonstrating the distribution of diatom taxa across oligotrophic to eutrophic preference classes for a representative sample.

14.8 MAKING FULL VAN DAM MULTI-INDEX PLOT IN GGPLOT2 (WITH STACKED BARS)

```
library(ggplot2)
library(ggthemes)

# Optional: Rename 'Indet' to 'Unclassified'
vd_long$Class <- ifelse(vd_long$Class == "Indet", "Unclassified", vd_long$Class)

# Turn Class into factor with order (you can also rearrange manually)
vd_long$Class <- factor(vd_long$Class,
                        levels = c("1", "2", "3", "4", "5", "6", "7", "Unclassified"))

ggplot(vd_long, aes(x = Sample, y = Percentage, fill = Class)) +
  geom_col(width = 0.9) +
  facet_wrap(~ Trait, scales = "free_y", ncol = 3) +
  scale_fill_brewer(palette = "Set3", na.translate = FALSE) +
  labs(
    title = "Van Dam Ecological Preferences by Sample",
    x = "Sample",
    y = "Percentage (%)",
    fill = "Ecological Class"
  ) +
  theme_minimal(base_size = 18) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5, size = 20),
    axis.title = element_text(face = "bold", size = 18),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 14, face = "bold"),
    axis.text.y = element_text(size = 16),
    strip.text = element_text(size = 17, face = "bold"),
    legend.title = element_text(face = "bold", size = 16),
    legend.text = element_text(size = 15),
    legend.position = "bottom"
  )
#Refer Fig. 14.15
```

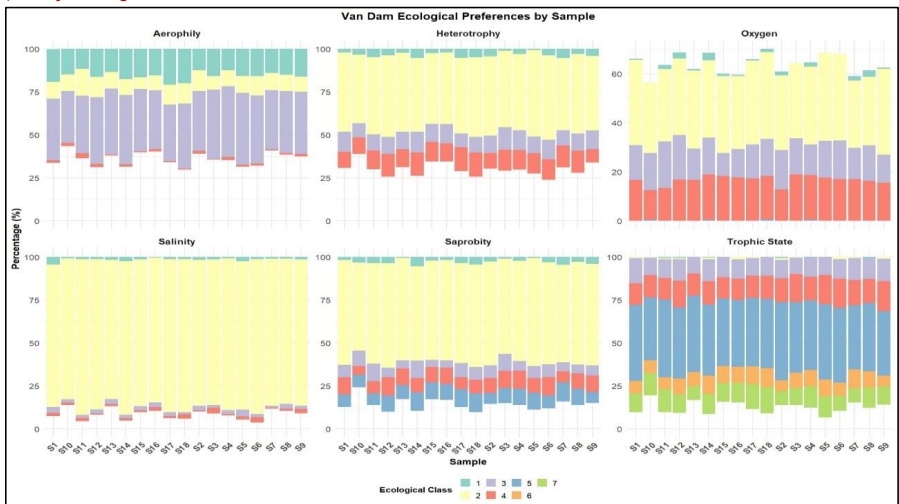


Fig. 14.14 Example of a trophic state plot derived from *DiaThor* outputs, demonstrating the distribution of diatom taxa across oligotrophic to eutrophic preference classes for a representative sample.

14.9 DESCRIPTIVE PLOTS FOR EACH VAN DAM INDEX USING A REUSABLE FUNCTION IN R

```
library(dplyr)
library(tidyr)
library(ggplot2)
```

Function to plot Van Dam index with descriptive labels

```
plot_vandam_index <- function(data, index_prefix, label_list, title_text) {
```

Select relevant columns

```
  relevant_cols <- grep(paste0("^", index_prefix, "\\.(\\d|Indet)$"), names(data), value
= TRUE)
  data_long <- data %>%
    mutate(Sample = rownames(.)) %>%
    select(Sample, all_of(relevant_cols)) %>%
    pivot_longer(-Sample, names_to = "Class", values_to = "Percentage") %>%
    mutate(Class = gsub(paste0(index_prefix, "\\."), "", Class)) %>%
    mutate(Class = factor(Class, levels = names(label_list), labels =
unname(label_list)))
```

Plot

```
  ggplot(data_long, aes(x = Sample, y = Percentage, fill = Class)) +
    geom_bar(stat = "identity") +
    theme_minimal() +
    labs(title = title_text, x = "Sample", y = "Percentage (%)") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
          axis.title = element_text(face = "bold"),
          plot.title = element_text(face = "bold", size = 16),
          legend.title = element_text(face = "bold")) +
    scale_fill_brewer(palette = "Paired")
}
```

For Oxygen Requirements, run:

```
plot_vandam_index(
  data = results_vandam,
  index_prefix = "VD.Oxygen",
  label_list = c("1" = "Aerophilic", "2" = "Oxyphilous", "3" = "Moderate oxygen",
                 "4" = "Low oxygen", "5" = "Anoxic", "Indet" = "Indeterminate"),
  title_text = "Van Dam Oxygen Requirements"
)
```

You can loop this for all 6 indices:

```
vandam_indices <- list(
  "VD.Salinity" = list(title = "Salinity Preferences", labels = c("1"="Fresh-brackish",
"2"="Freshwater", "3"="Brackish", "4"="Marine", "Indet"="Indeterminate")),
  "VD.N.Het" = list(title = "Nitrogen Heterotrophy", labels = c("1"="Obligate
autotrophs", "2"="Facultative heterotrophs", "3"="Heterotrophs",
"4"="Photoheterotrophs", "Indet"="Indeterminate")),
  "VD.Oxygen" = list(title = "Oxygen Requirements", labels = c("1"="Aerophilic",
"2"="Oxyphilous", "3"="Moderate oxygen", "4"="Low oxygen", "5"="Anoxic",
"Indet"="Indeterminate")),
  "VD.Saprobity" = list(title = "Saprobity Classes", labels = c("1"="Oligosaprobic",
"2"="β-mesosaprobic", "3"="α-mesosaprobic", "4"="Polysaprobic", "5"="Hypertrophic",
"Indet"="Indeterminate")),
```

```

"VD.Aero" = list(title = "Aerophily", labels = c("1"="Subaerial", "2"="Air-water
interface", "3"="Aquatic", "4"="Low oxygen aquatic", "5"="Anaerobic",
"Indet"="Indeterminate")),
"VD.Trophic" = list(title = "Trophic State", labels = c("1"="Oligotraphentic",
"2"="Oligo-mesotraphentic", "3"="Mesotraphentic", "4"="Meso-eutraphentic",
"5"="Eutraphentic", "6"="Hyper-eutraphentic", "7"="Polysaprobic",
"Indet"="Indeterminate"))
)

```

Loop through all: It will generate 6 graphs for all the traits with descriptive labels

```

for (index in names(vandam_indices)) {
  cat("Plotting", vandam_indices[[index]]$title, "\n")
  print(
    plot_vandam_index(
      data = results_vandam,
      index_prefix = index,
      label_list = vandam_indices[[index]]$labels,
      title_text = paste("Van Dam", vandam_indices[[index]]$title)
    )
  )
} #Refer Fig. 14.16-14.19

```

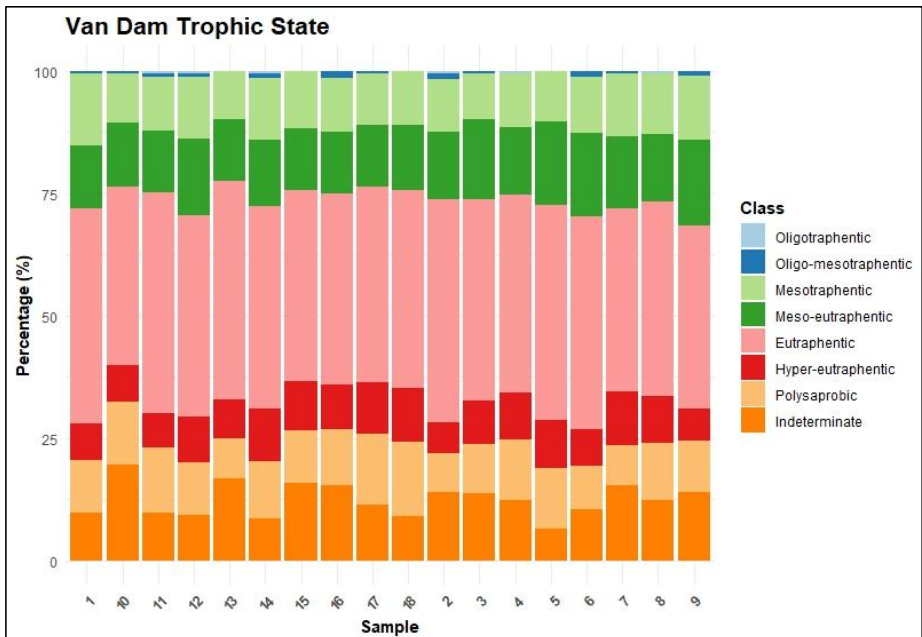


Fig.14.16 Example of a trophic state plot derived from DiaThor outputs, demonstrating the distribution of diatom taxa across oligotrophic to eutrophic preference classes for a representative sample.

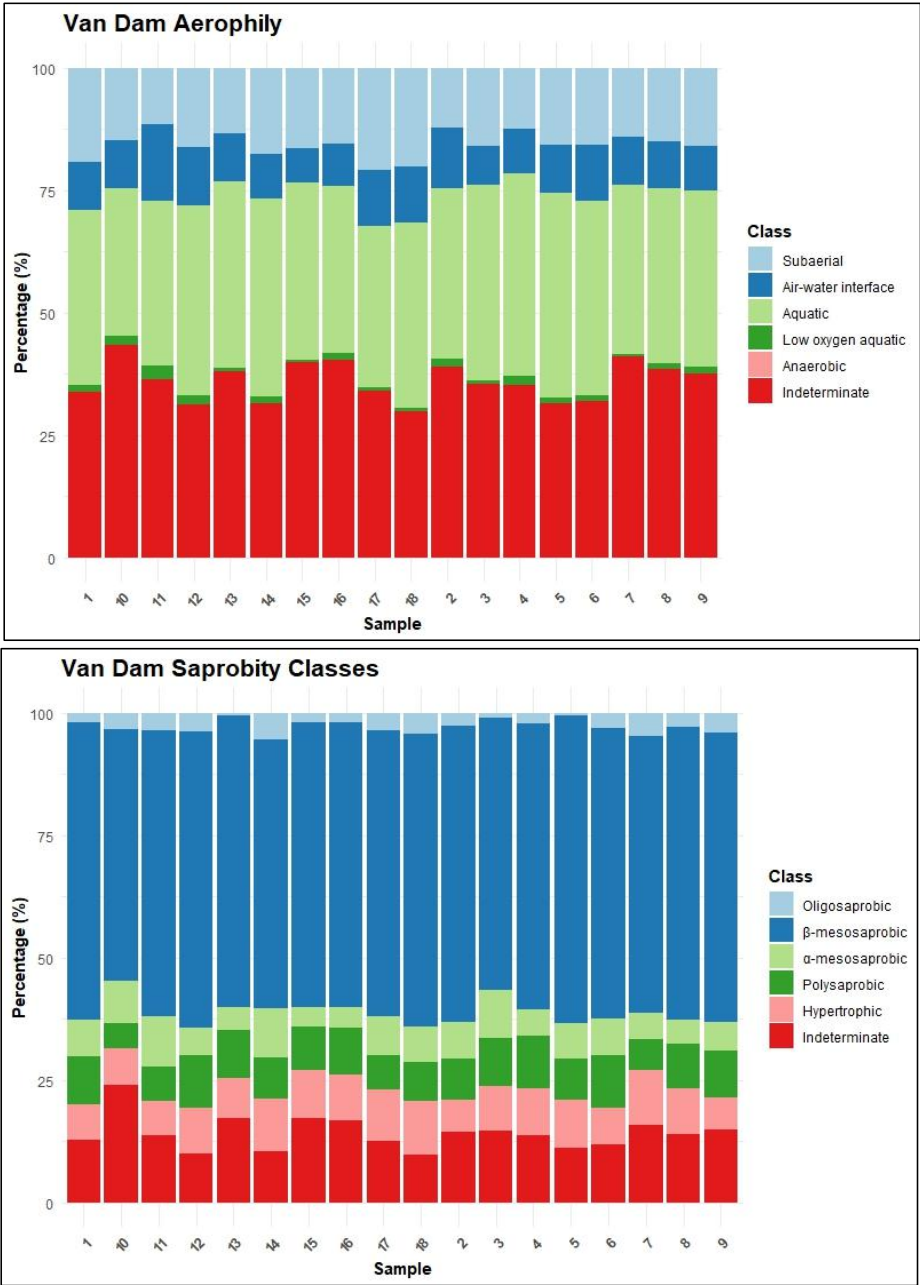


Fig. 14.17 Example of plots derived from DiaThor outputs, demonstrating the distribution of diatom taxa across oligotrophic to eutrophic preference classes for a representative sample.

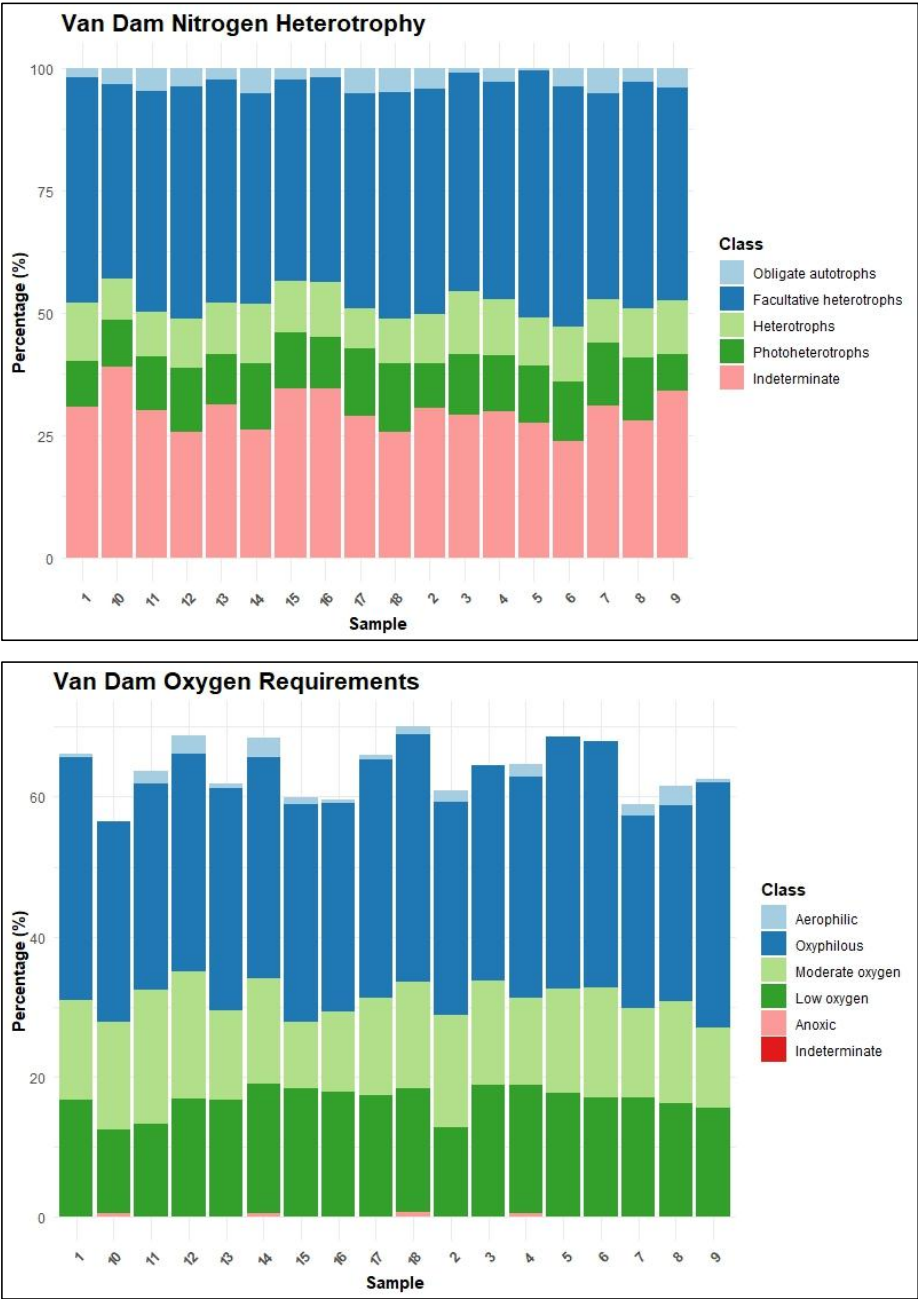


Fig. 14.18 Example of plots derived from DiaThor outputs, demonstrating the distribution of diatom taxa across oligotrophic to eutrophic preference classes for a representative sample.

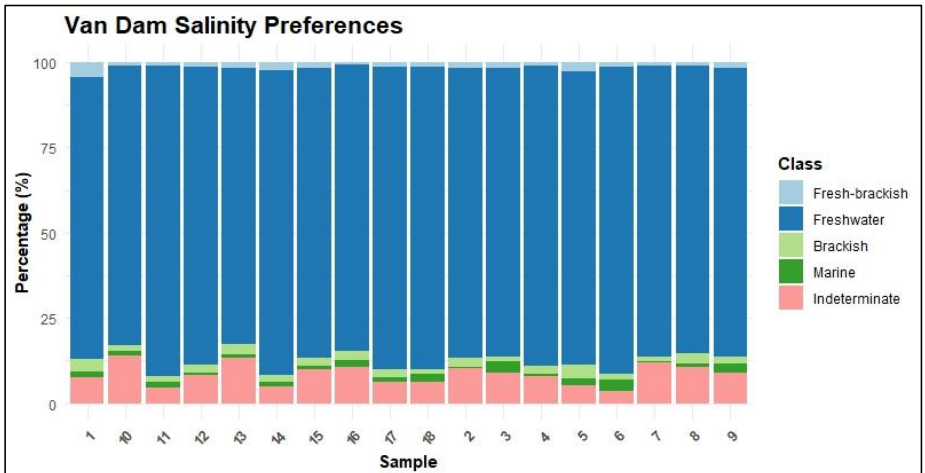


Fig. 14.19 Example of plots derived from *DiaThor* outputs, demonstrating the distribution of diatom taxa across different preference classes for a representative sample.

#EXPLANATION

```
plot_vandam_index()
plot_vandam_index <- function(data, index_prefix, label_list, title_text) {Define a function
  with:
  data = your results_vandam dataframe index_prefix = prefix of column names (like "VD.Oxygen")
  label_list = named vector for descriptive labels title_text = plot title (string)
  relevant_cols <- grep(paste0("^", index_prefix, "\\.(\\d|Indet)$"), names(data), value
    = TRUE)
  Searches column names that match the pattern like VD.Oxygen.1, VD.Oxygen.Indet using grep(). Only includes
  numbered or 'Indet' classes. For example, if index_prefix = "VD.Oxygen", this line will grab:
  c("VD.Oxygen.1", "VD.Oxygen.2", ..., "VD.Oxygen.5", "VD.Oxygen.Indet")
  data_long <- data %>%
    mutate(Sample = rownames(.)) %>%
    select(Sample, all_of(relevant_cols)) %>%
  pivot_longer(-Sample, names_to = "Class", values_to = "Percentage") %>%
    mutate(Class = gsub(paste0(index_prefix, "\\."), "", Class)) %>%
  mutate(Class = factor(Class, levels = names(label_list), labels = unname(label_list)))
  This line does the transformation magic
  mutate(Sample = rownames(.))
  Adds a 'Sample' column from rownames if not already present.
  select(Sample, all_of(relevant_cols))
  Keeps only the Sample column and the relevant Van Dam columns.
  pivot_longer()
  Converts the data from wide to long format, which ggplot2 prefers:
  Original columns like VD.Oxygen.1, 2, 3... become values under a single column Class.
  mutate(Class = gsub(...))
  Removes the prefix (VD.Oxygen.) to leave only the class label (1, 2, Indet...).
  mutate(Class = factor(...))
  Turns Class into a factor with custom labels defined by you. This is how it gets into readable format
  ggplot(data_long, aes(x = Sample, y = Percentage, fill = Class)) +
    geom_bar(stat = "identity") +
    theme_minimal() +
    labs(title = title_text, x = "Sample", y = "Percentage (%)") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
```



```
axis.title = element_text(face = "bold"),
plot.title = element_text(face = "bold", size = 16),
legend.title = element_text(face = "bold")) +
scale_fill_brewer(palette = "Paired")
```

This builds the bar plot:

```
aes(x = Sample, y = Percentage, fill = Class) sets the aesthetics.
```

```
geom_bar(stat = "identity") builds a stacked bar chart.
```

```
theme_minimal() keeps the plot clean.
```

```
labs() gives axis labels and plot title.
```

```
theme() tweaks styling:
```

```
Bold x-axis labels (rotated 45°), Bold titles and legend.
```

```
scale_fill_brewer() applies a color palette from RColorBrewer for visual clarity.
```

```
Loop- To plot all indices in one go
```

```
vandam_indices <- list(
```

```
"VD.Salinity" = list(title = "Salinity Preferences", labels = c("1"="Fresh-brackish",
...)), ...)
```

This creates a named list of Van Dam indices you want to plot: The name (e.g. 'VD.Salinity') is the prefix.

Each list entry has: title: for the plot title; labels: descriptive labels for classes like '1'='Fresh-brackish'

```
for (index in names(vandam_indices)) {
  cat("Plotting", vandam_indices[[index]]$title, "\n")
  print(
    plot_vandam_index(
      data = results_vandam,
      index_prefix = index,
      label_list = vandam_indices[[index]]$labels,
      title_text = paste("Van Dam", vandam_indices[[index]]$title)
    )
  )
}
```

The loop goes through each ecological index in the list. Calls your plot_vandam_index() function for each. Prints a message to the console, and displays the plot. This setup: Automates plotting for all Van Dam indices; gives you labelled, and consistent visuals. Supports further expansion (like saving to PNG, arranging in grids, adding tooltips with plotly, etc.).

14.10 MULTIPANEL PLOT GRID STYLE (2 X 2)**CONSTRUCTING MULTIPANEL ANALYTICAL GRIDS (PATCHWORK & COWPLOT) (VAN DAM TRAITS)**

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(patchwork) # For method 1
library(cowplot)   # For method 2
```

Example labels (add more as needed)

```
labels_salinity <- c("1"="Fresh-brackish", "2"="Fresh", "3"="Brackish", "4"="Marine",
"Indet"="Indeterminate")
labels_oxygen   <- c("1"="Aerophil", "2"="High", "3"="Moderate", "4"="Low", "5"="Anoxic",
"Indet"="Indeterminate")
labels_trophic  <- c("1"="Oligo", "2"="Oligo-Meso", "3"="Meso", "4"="Meso-Eu",
"5"="Eutrophic", "6"="Hypereu", "7"="Dystrophic", "Indet"="Indeterminate")
labels_saprobity <- c("1"="Oligosaprobic", "2"="β-Mesosaprobic", "3"="α-Mesosaprobic",
"4"="Polysaprobic", "5"="Hyperpolysaprobic", "Indet"="Indeterminate")
```

Generate the plots

```
p_salinity <- plot_vandam_index(results_vandam, "VD.Salinity", labels_salinity, "Van Dam
- Salinity")
p_oxygen   <- plot_vandam_index(results_vandam, "VD.Oxygen", labels_oxygen, "Van Dam -
Oxygen")
p_trophic  <- plot_vandam_index(results_vandam, "VD.Trophic", labels_trophic, "Van Dam -
Trophic State")
p_saprobity <- plot_vandam_index(results_vandam, "VD.Saprobity", labels_saprobity, "Van
Dam - Saprobity")
```

Arrange them in a 2x2 grid

```
(p_salinity | p_oxygen) /
(p_trophic  | p_saprobity)
```

Save the plot in png format

```
ggsave("VanDam_Multipanel_Patchwork.png", width = 16, height = 10, dpi = 300)
```

Arrange and label

```
multi_plot <- plot_grid(
  p_salinity, p_oxygen, p_trophic, p_saprobity,
  labels = c("A", "B", "C", "D"),
  ncol = 2, align = "hv"
)
```

Show it

```
print(multi_plot)
```

Save it

```
ggsave("VanDam_Multipanel_Cowplot.png", plot = multi_plot, width = 16, height = 10, dpi
= 300) ggsave("VanDam_Multipanel.pdf", plot = multi_plot, width = 16, height = 10)
```

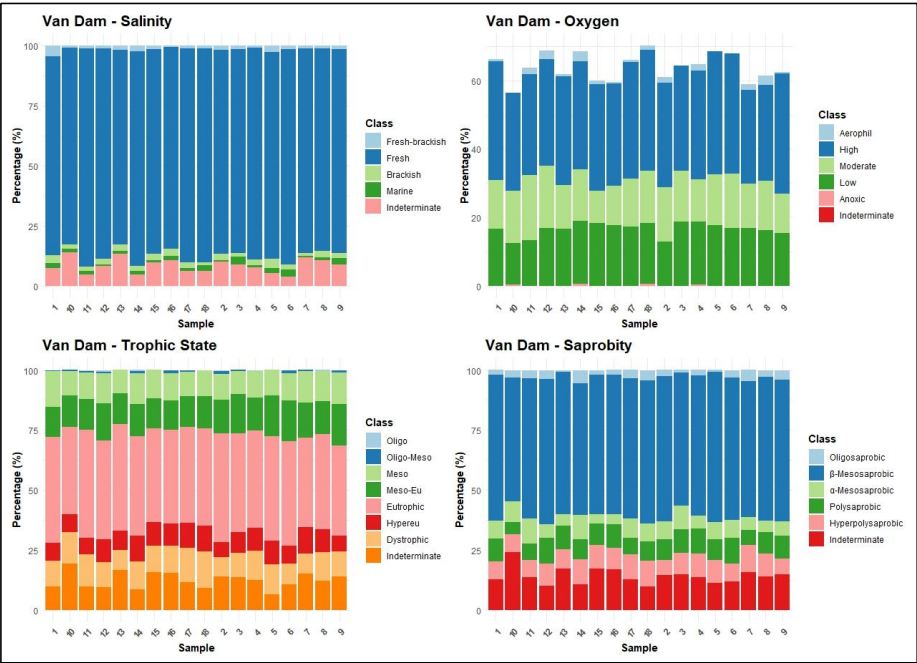


Fig.14.20 Multipanel plot arranged in a 2×2 grid, visualizing selected Van Dam ecological traits simultaneously to enable comparative interpretation of diatom-based environmental preferences across samples.

14.11 MULTIPANEL PLOT GRID STYLE (3 X 2)

CONSTRUCTING MULTIPANEL ANALYTICAL GRIDS (PATCHWORK & COWPLOT) (VAN DAM TRAITS)

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(patchwork) # For method 1
library(cowplot)   # For method 2

plot_vandam_index <- function(data, index_prefix, class_labels, title) {
  index_cols <- grep(paste0("^", index_prefix, "\\."), names(data), value = TRUE)
  df_long <- data %>%
    mutate(Sample = paste0("S", row_number())) %>%
    select(Sample, all_of(index_cols)) %>%
    pivot_longer(cols = -Sample, names_to = "Class", values_to = "Percentage") %>%
    mutate(
      Class = gsub(paste0(index_prefix, "\\."), "", Class),
      Class = factor(Class, levels = names(class_labels), labels = class_labels)
    )

  ggplot(df_long, aes(x = Sample, y = Percentage, fill = Class)) +
    geom_bar(stat = "identity") +
    labs(title = title, x = "Sample", y = "Percentage (%)", fill = "Class") +
    theme_minimal(base_size = 12) +
    theme(
      axis.text.x = element_text(angle = 45, hjust = 1),
      plot.title = element_text(face = "bold", size = 14),
      axis.title = element_text(face = "bold")
    )
}
```

#Define Labels for Each Index

```
# ◇ 1. Aerophily (Oxygen exposure)
labels_aero <- c(
  "1" = "Aerophilic",
  "2" = "High oxygen",
  "3" = "Moderate",
  "4" = "Low oxygen",
  "5" = "Anoxic",
  "Indet" = "Indeterminate"
)

# ◇ 2. Heterotrophy (Nutritional mode)
labels_hetero <- c(
  "1" = "Obligate autotrophs",
  "2" = "Facultative autotrophs",
  "3" = "Facultative heterotrophs",
  "4" = "Obligate heterotrophs",
  "Indet" = "Indeterminate"
)

# ◇ 3. Oxygen Requirements
labels_oxygen <- c(
  "1" = "Aerophilic",
  "2" = "Oxyphilic",
```

```

    "3" = "Moderate",
    "4" = "Tolerant",
    "5" = "Anoxic",
    "Indet" = "Indeterminate"
  )

# ◊ 4. Salinity
labels_salinity <- c(
  "1" = "Fresh-brackish",
  "2" = "Freshwater",
  "3" = "Brackish",
  "4" = "Marine",
  "Indet" = "Indeterminate"
)

# ◊ 5. Saprobity (Organic pollution tolerance)
labels_saprobity <- c(
  "1" = "Oligosaprobic",
  "2" = "β-Mesosaprobic",
  "3" = "α-Mesosaprobic",
  "4" = "Polysaprobic",
  "5" = "Hypersaprobic",
  "Indet" = "Indeterminate"
)

# ◊ 6. Trophic State (Nutrient preference)
labels_trophic <- c(
  "1" = "Oligotrophic",
  "2" = "Oligo-Mesotrophic",
  "3" = "Mesotrophic",
  "4" = "Meso-Eutrophic",
  "5" = "Eutrophic",
  "6" = "Hypertrophic",
  "7" = "Dystrophic",
  "Indet" = "Indeterminate"
)

#Generate All Six Plots
p_aero <- plot_vandam_index(results_vandam, "VD.Aero", labels_aero, "Van Dam -
Aerophily")
p_hetero <- plot_vandam_index(results_vandam, "VD.N.Het", labels_hetero, "Van Dam -
Heterotrophy")
p_oxygen <- plot_vandam_index(results_vandam, "VD.Oxygen", labels_oxygen, "Van Dam -
Oxygen Requirements")
p_salinity <- plot_vandam_index(results_vandam, "VD.Salinity", labels_salinity, "Van Dam
- Salinity")
p_saprobity <- plot_vandam_index(results_vandam, "VD.Saprobity", labels_saprobity, "Van
Dam - Saprobity")
p_trophic <- plot_vandam_index(results_vandam, "VD.Trophic", labels_trophic, "Van Dam
- Trophic State")

# Arrange into 3x2 Grid with {patchwork}
library(patchwork)

final_plot <- (p_aero | p_hetero | p_oxygen) /
               (p_salinity | p_saprobity | p_trophic)

```

Display the full plot

```
print(final_plot)
```

Save the full plot in

i. PNG format

```
ggsave("VanDam_Ecological_Preferences_All.png", plot = final_plot, width = 18, height = 12, dpi = 300)
```

OR

ii. PDF format

```
ggsave("VanDam_Ecological_Preferences_All.pdf", plot = final_plot, width = 18, height = 12)
```

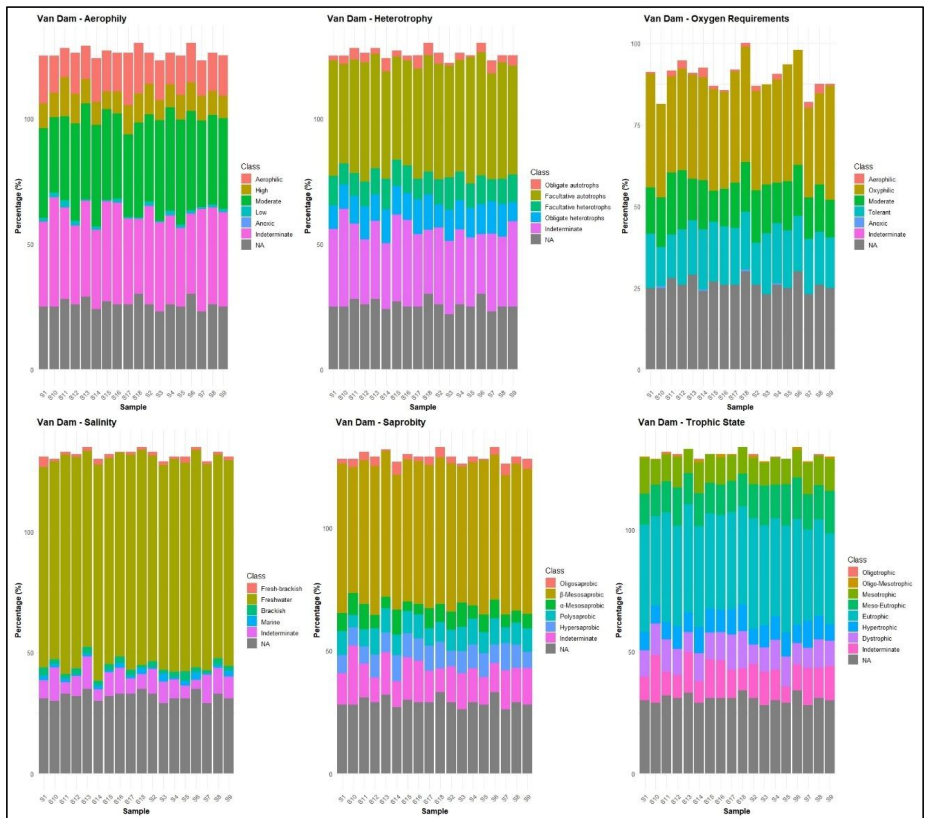


Fig. 14.21 Multipanel plot arranged in a 3×2 grid, visualizing selected Van Dam ecological traits simultaneously to enable comparative interpretation of diatom-based environmental preferences across samples.

14.12 CUSTOM FUNCTION PLOT_ALL_VANDAM()

Load your data and calculate VanDam index as you have studied in previous chapters. Then, copy paste this code.

```
plot_all_vandam <- function(results_vandam, output_basename =
"VanDam_Ecological_Preferences") {
  library(ggplot2)
  library(dplyr)
  library(tidyr)
  library(patchwork)

  # Inner function to generate one index plot
  plot_vandam_index <- function(data, index_prefix, class_labels, title) {
    index_cols <- grep(paste0("^", index_prefix, "\\."), names(data), value = TRUE)

    df_long <- data %>%
      mutate(Sample = paste0("S", row_number())) %>%
      select(Sample, all_of(index_cols)) %>%
      pivot_longer(cols = -Sample, names_to = "Class", values_to = "Percentage") %>%
      mutate(
        Class = gsub(paste0(index_prefix, "\\."), "", Class),
        Class = factor(Class, levels = names(class_labels), labels = class_labels)
      )

    ggplot(df_long, aes(x = Sample, y = Percentage, fill = Class)) +
      geom_bar(stat = "identity") +
      labs(title = title, x = "Sample", y = "Percentage (%)", fill = "Class") +
      scale_fill_brewer(palette = "Set2") +
      theme_minimal(base_size = 12, base_family = "serif") +
      theme(
        axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(face = "bold", size = 14),
        axis.title = element_text(face = "bold"),
        legend.position = "bottom"
      )
  }

  # Trait Labels
  labels_aero <- c("1" = "Aerophilic", "2" = "High oxygen", "3" = "Moderate", "4" = "Low
oxygen", "5" = "Anoxic", "Indet" = "Indeterminate")
  labels_hetero <- c("1" = "Obligate autotrophs", "2" = "Facultative autotrophs", "3" =
"Facultative heterotrophs", "4" = "Obligate heterotrophs", "Indet" = "Indeterminate")
  labels_oxygen <- c("1" = "Aerophilic", "2" = "Oxyphilic", "3" = "Moderate", "4" =
"Tolerant", "5" = "Anoxic", "Indet" = "Indeterminate")
  labels_salinity <- c("1" = "Fresh-brackish", "2" = "Freshwater", "3" = "Brackish", "4"
= "Marine", "Indet" = "Indeterminate")
  labels_saprobity <- c("1" = "Oligosaprobic", "2" = " $\beta$ -Mesosaprobic", "3" = " $\alpha$ -
Mesosaprobic", "4" = "Polysaprobic", "5" = "Hypersaprobic", "Indet" = "Indeterminate")
  labels_trophic <- c("1" = "Oligotrophic", "2" = "Oligo-Mesotrophic", "3" =
"Mesotrophic", "4" = "Meso-Eutrophic", "5" = "Eutrophic", "6" = "Hypertrophic", "7" =
"Dystrophic", "Indet" = "Indeterminate")

  # Generate Each Plot
  p1 <- plot_vandam_index(results_vandam, "VD.Aero", labels_aero, "Van Dam - Aerophily")
  p2 <- plot_vandam_index(results_vandam, "VD.N.Het", labels_hetero, "Van Dam -
Heterotrophy")
  p3 <- plot_vandam_index(results_vandam, "VD.Oxygen", labels_oxygen, "Van Dam - Oxygen
Requirements")
}
```



```
p4 <- plot_vandam_index(results_vandam, "VD.Salinity", labels_salinity, "Van Dam - Salinity")
p5 <- plot_vandam_index(results_vandam, "VD.Saprobity", labels_saprobity, "Van Dam - Saprobity")
p6 <- plot_vandam_index(results_vandam, "VD.Trophic", labels_trophic, "Van Dam - Trophic State")
```

Combine in 3x2 grid

```
final_plot <- (p1 | p2 | p3) / (p4 | p5 | p6) +
  plot_layout(guides = "collect") +
  plot_annotation(title = "Van Dam Ecological Preferences", theme = theme(plot.title = element_text(size = 16, face = "bold")))
```

Show it

```
print(final_plot)
```

Save

```
ggsave(paste0(output_basename, ".png"), plot = final_plot, width = 18, height = 12, dpi = 300)
ggsave(paste0(output_basename, ".pdf"), plot = final_plot, width = 18, height = 12)
}
```

Now a custom function code has been written, use it:

```
plot_all_vandam(results_vandam)
```

You can also rename the file if you want:

```
plot_all_vandam(results_vandam, output_basename = "Ecological_Traits_Plot")
```

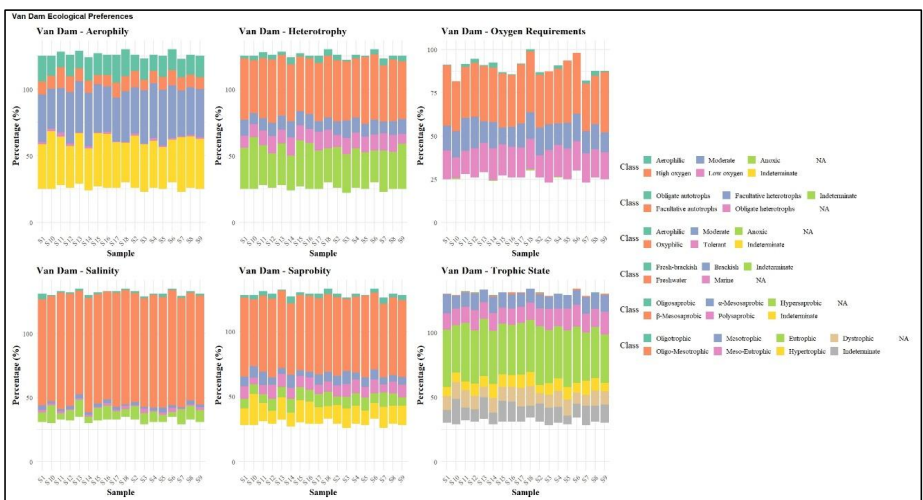


Fig. 14.22 Output of the custom function `plot_all_vandam()`, generating a consolidated set of Van Dam ecological trait plots to provide an integrated visualization of diatom-based environmental preferences across samples.

14.13 DESCRIPTIVE LABELS FOR VANDAM TRAIT PLOTS

1. Aerophily (Oxygen exposure)

```
labels_aero <- c(
  "1" = "Aerophilic",
  "2" = "High oxygen",
  "3" = "Moderate",
  "4" = "Low oxygen",
  "5" = "Anoxic",
  "Indet" = "Indeterminate"
)
```

2. Heterotrophy (Nutritional mode)

```
labels_hetero <- c(
  "1" = "Obligate autotrophs",
  "2" = "Facultative autotrophs",
  "3" = "Facultative heterotrophs",
  "4" = "Obligate heterotrophs",
  "Indet" = "Indeterminate"
)
```

3. Oxygen Requirements

```
labels_oxygen <- c(
  "1" = "Aerophilic",
  "2" = "Oxyphilic",
  "3" = "Moderate",
  "4" = "Tolerant",
  "5" = "Anoxic",
  "Indet" = "Indeterminate"
)
```

4. Salinity

```
labels_salinity <- c(
  "1" = "Fresh-brackish",
  "2" = "Freshwater",
  "3" = "Brackish",
  "4" = "Marine",
  "Indet" = "Indeterminate"
)
```

5. Saprobity (Organic pollution tolerance)

```
labels_saprobity <- c(
  "1" = "Oligosaprobic",
  "2" = " $\beta$ -Mesosaprobic",
  "3" = " $\alpha$ -Mesosaprobic",
  "4" = "Polysaprobic",
  "5" = "Hypersaprobic",
  "Indet" = "Indeterminate"
)
```

6. Trophic State (Nutrient preference)

```
labels_trophic <- c(
  "1" = "Oligotrophic",
  "2" = "Oligo-Mesotrophic",
  "3" = "Mesotrophic",
  "4" = "Meso-Eutrophic",
  "5" = "Eutrophic",
  "6" = "Hypertrophic",
  "7" = "Dystrophic",
  "Indet" = "Indeterminate"
)
```

7. SPEAR (Sensitivity to herbicides)

```
labels_spear <- c(
  "0" = "SPEAR-insensitive",
  "1" = "SPEAR-sensitive"
)
```

8. SLA (Saprobic levels by Sládeček)

```
labels_sla <- c(
  "1" = "Oligosaprobic",
  "2" = " $\beta$  -Mesosaprobic",
  "3" = " $\alpha$  -Mesosaprobic",
  "4" = "Polysaprobic",
  "Indet" = "Indeterminate"
)
```

9. IDAP (Agricultural pressure)

```
labels_idap <- c(
  "1" = "Very low pressure",
  "2" = "Low pressure",
  "3" = "Moderate pressure",
  "4" = "High pressure",
  "Indet" = "Indeterminate"
)
```

10. Guilds (Functional ecology)

```
labels_guilds <- c(
  "H" = "High profile",
  "L" = "Low profile",
  "M" = "Motile",
  "P" = "Planktonic",
  "Indet" = "Indeterminate"
)
```

14.14 MULTI-INDEX PLOTTING FUNCTION TO VISUALIZE ANY VAN DAM ECOLOGICAL TRAIT

Create a Generic Plot Function for Any Van Dam Trait

```
library(tidyverse)

plot_vandam_trait <- function(data, trait_prefix, plot_type = "area") {

  # Step 1: Filter trait columns + Sample
  df_long <- data %>%
    select(Sample, starts_with(trait_prefix)) %>%
    pivot_longer(cols = -Sample, names_to = "Class", values_to = "Percentage")

  # Clean up class labels (remove prefix)
  df_long$Class <- str_remove(df_long$Class, paste0("^", trait_prefix, "\\."))

  # Step 2: Plot based on type
  p <- ggplot(df_long, aes(x = Sample, y = Percentage, fill = Class, group = Class))

  if (plot_type == "area") {
    p <- p + geom_area(position = "stack", alpha = 0.8)
  } else if (plot_type == "line") {
    p <- p + geom_line(aes(color = Class), size = 1) + geom_point(size = 2)
  } else if (plot_type == "heatmap") {
    p <- ggplot(df_long, aes(x = Sample, y = Class, fill = Percentage)) +
      geom_tile(color = "white") +
      scale_fill_viridis_c()
  } else if (plot_type == "dot") {
    p <- ggplot(df_long, aes(x = Sample, y = Class, size = Percentage, color = Percentage))
  }
  +
    geom_point(alpha = 0.7) +
    scale_color_viridis_c() +
    scale_size(range = c(2, 10))
  } else {
    stop("Unsupported plot type. Choose from 'area', 'line', 'heatmap', or 'dot'")
  }

  # Add styling
  p <- p +
    theme_minimal(base_size = 14) +
    labs(title = paste("Van Dam -", trait_prefix, "Classes"),
         x = "Sample", y = "Percentage (%)") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
          axis.title = element_text(face = "bold"),
          plot.title = element_text(face = "bold", size = 16))

  return(p)
}
```

Use the Function:

Plot Oxygen Classes (stacked area)

```
plot_vandam_trait(results_vandam, "VD.Oxygen", "area")
```

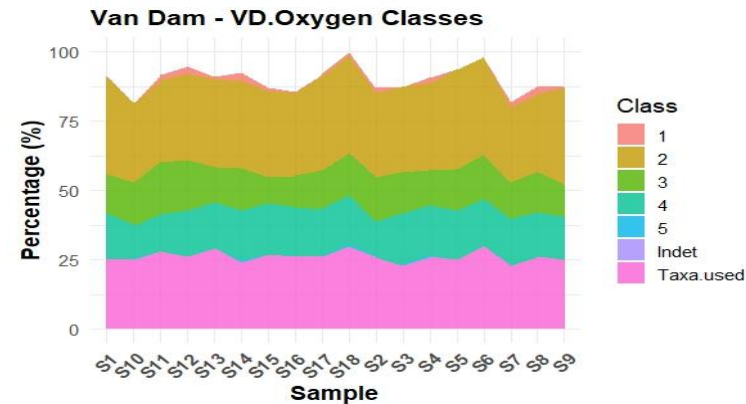


Fig. 14.23 Stacked area plot illustrating the distribution of oxygen requirement classes across samples, highlighting shifts in diatom community composition and inferred oxygen conditions derived from DiaThor outputs.

Plot Trophic Classes (line plot)

```
plot_vandam_trait(results_vandam, "VD.Trophic", "line")
```

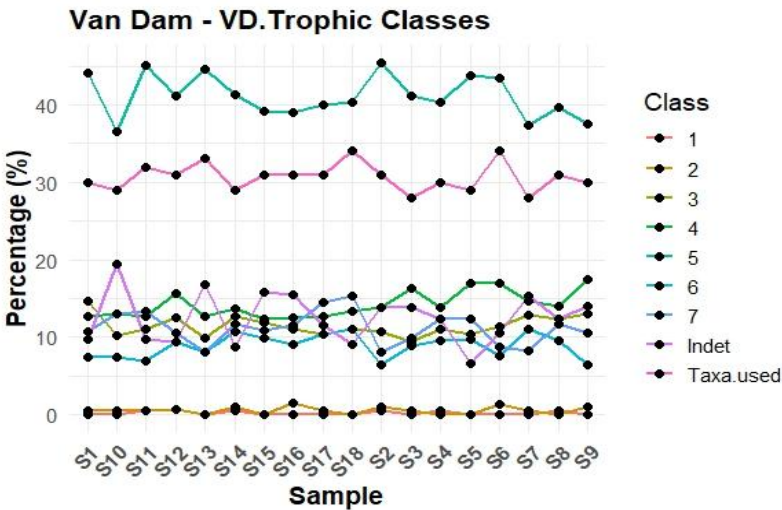


Fig. 14.24 Line plot depicting trends in trophic state classes across samples, illustrating changes in nutrient preference and ecosystem productivity inferred from DiaThor outputs.

Plot Aerophily (dot plot)

```
plot_vandam_trait(results_vandam, "VD.Aero", "dot")
```

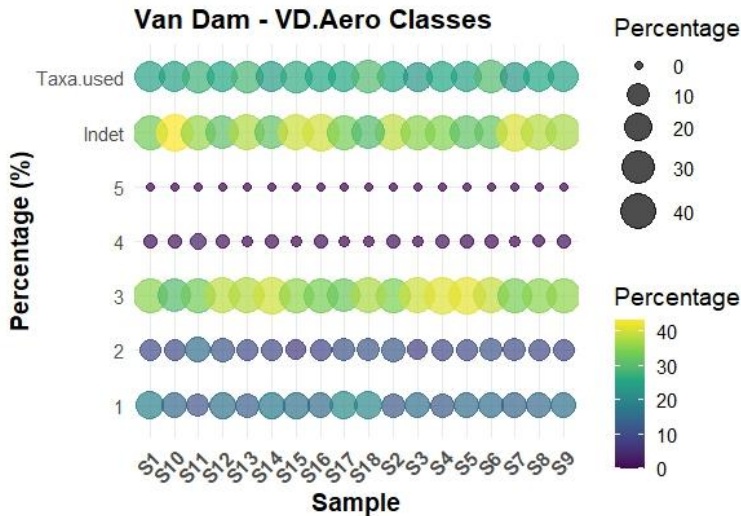


Fig. 14.25 Dot plot showing the distribution of aerophily classes across samples, highlighting differences in diatom tolerance to aerial exposure and hydrological variability as inferred from DiaThor outputs.

Plot Heterotrophy (heatmap)

```
plot_vandam_trait(results_vandam, "VD.N.Het", "heatmap")
```

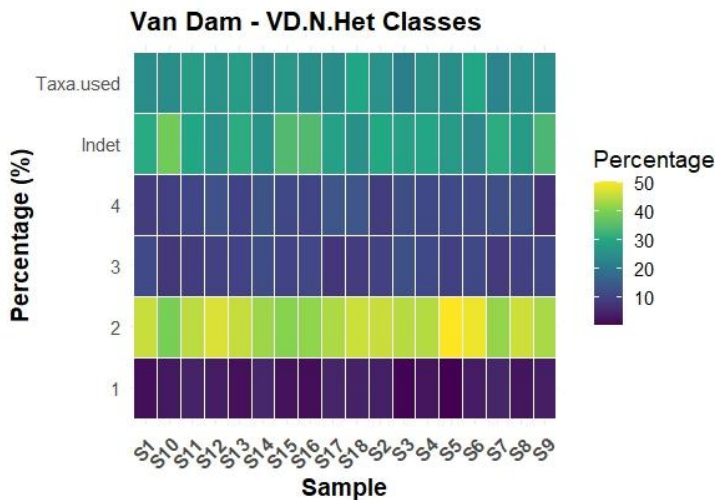


Fig. 14.26 Heatmap illustrating the distribution of heterotrophy classes across samples, revealing patterns in nitrogen utilization strategies and ecological gradients inferred from DiaThor outputs through color intensity and clustering

CHAPTER 16

PARAMETERS FOR INTERPRETING THE RESULT OF
VARIOUS CALCULATED INDICES

*"Numbers alone don't speak the truth – interpretation gives them a voice,
and parameters give them meaning."*

While computational tools can swiftly generate a wide array of diatom-based metrics, raw numerical outputs require proper contextualization to be scientifically meaningful. To prevent the core objective of running a single index from being obscured by overwhelming code or technical execution details, this chapter abstracts the programming mechanics. Instead, it focuses exclusively on the ecological translation of the results.

Ecological indices are not universal; they utilize distinct numerical scales—ranging from 1 to 5, 0 to 100, or 1 to 20—depending on their original geographic calibration and the specific environmental stressor they target (e.g., organic pollution, nutrient enrichment, or herbicide impact). Furthermore, trait-based metrics like ecological guilds and size classes rely on percentage compositions rather than fixed scores. By applying the standardized parameter bands detailed in the framework below, researchers can reliably translate these computational scores into actionable ecological insights, assessing everything from trophic status to habitat resilience. Once ecological indices have been calculated, the resulting numerical values must be translated into categorical quality classes to be useful for environmental monitoring and policy management. This chapter outlines the standard interpretation parameters for various diatom-based indices. The reference scales provided below classify water bodies into ecological status categories (e.g., High, Good, Moderate, Poor, Bad) based on established European and regional frameworks.

Function	Index Name	Value Range	Interpretation Bands	What It Tells Us
<code>diat_ips()</code>	Pollution Sensitivity Index (IPS)	1–20	>15 = High; 12–15 = Good; 9–12 = Moderate; <9 = Poor	Sensitivity to organic and general pollution
<code>diat_tdi()</code>	Trophic Diatom Index (TDI)	0–100	<30 = Oligotrophic; 30–50 = Mesotrophic; >50 = Eutrophic	Nutrient levels (esp. phosphorus); eutrophication indicator
<code>diat_sla()</code>	Sládeček Index	1–5	1–2 = Clean; 3 = Moderate; 4–5 = Polluted	Organic matter and wastewater pollution
<code>diat_lobo()</code>	LOBO Index (Brazil)	1–5	1 = Clean; 5 = Polluted	General ecological condition in tropical systems
<code>diat_edi()</code>	Ecological Diatom Index	0–5	>4 = Very Good; 3–4 = Good; 2–3 = Moderate; <2 = Bad	Overall ecological quality, especially in Mediterranean rivers

diat_ilm()	Italian Diatom Index	1–18	>15 = High; 12–15 = Good; 9–12 = Moderate; <9 = Poor	Pollution sensitivity in Italian river systems
diat_idch()	Chilean Diatom Index	1–5	1–2 = Good; 3 = Moderate; 4–5 = Polluted	Chilean rivers water quality
diat_ddi()	Descy Diatom Index	0–5	>4 = Clean; 3 = Moderate; <2 = Polluted	Saprobity and general water quality
diat_epid()	EPID Index	1–5	1 = Clean; 5 = Polluted	Health of epiphytic communities
diat_idp()	Pedological Diatom Index	Varies	High = Natural soils; Low = Disturbed/urban soils	Soil biological quality
diat_idap()	Diatom Soil Activity Index	0–1	>0.7 = High activity; <0.3 = Low activity	Biological activity and human impact on soil
diat_cemfgs_rb()	Brazilian Regional Index	Varies	Region-specific bands	Water quality based on local species
diat_spear()	SPEARherbicides	0–1	>0.6 = Low herbicide stress; <0.4 = High stress	Herbicide impact on sensitive diatoms
diat_pbidw()	Pamela Biotic Index (Weighted)	Experimental	To be determined	Under research
diat_vandam()	Van Dam Ecological Preferences	Category-based	Saprobity: 1–4; Salinity: 1–5; Trophic: 1–5	Species' ecological preferences
diat_guilds()	Ecological Guilds	% composition	Dominance of motile/planktonic = disturbance	Community structure and disturbance response
diat_morpho()	Morphological Groups	% composition	High low-profile taxa = clean; motile = disturbed	Functional traits (mobility, shape, structure)

diat_size()	Size Classes	% composition	Large taxa decline in polluted sites	Ecosystem health and structure
diat_diversity()	Shannon, Simpson, Evenness	0–3 (Shannon)	>2.5 = High diversity; <1 = Low diversity	Biodiversity & ecosystem resilience
diat_disp()	Dispersion	0–1	High = community variability between sites	Beta diversity / habitat heterogeneity

Interpretation & Usefulness of DiaThor Indices Each index provides both interpretive bands and ecological insights	
Van Dam Ecological Preferences	Saprobity 1–4; Salinity 1–5; Trophic 1–5 Ecological preferences (traits)
Trophic Diatom Index (TDI)	<30 = Oligotrophic; 30–50 = Mesotrophic; >50 = Eutrophic Nutrient levels & eutrophication
SPEARherbicides	>0.6 = Low herbicide stress; <0.4 = High stress Herbicide stress on diatoms
Sládeček Index	1–2 = Clean; 3 = Moderate; 4–5 = Polluted Organic matter & wastewater pollution
Size Classes	Large taxa decline = polluted Ecosystem structure & resilience
Pollution Sensitivity Index (IPS)	>15 = High; 12–15 = Good; 9–12 = Moderate; <9 = Poor Pollution sensitivity (organic & general)
Pedological Diatom Index	High = Natural soils; Low = Disturbed soils Soil biological quality
Pamela Biotic Index (Weighted)	Experimental Experimental biotic index
Morphological Groups	Profiles vs motiles = clean vs disturbed Functional traits (mobility, shape)
LOBO Index (Brazil)	1 = Clean; 5 = Polluted Ecological quality (tropics)
Italian Diatom Index (ILM)	>15 = High; 12–15 = Good; 9–12 = Moderate; <9 = Poor Pollution sensitivity (Italy)
EPID Index	1 = Clean; 5 = Polluted Epiphytic community health
Ecological Guilds	% composition of guilds Community structure response
Ecological Diatom Index (EDI)	>4 = Very Good; 3–4 = Good; 2–3 = Moderate; <2 = Bad Overall ecological quality (rivers)
Diversity Indices	>2.5 = High diversity; <1 = Low diversity Biodiversity & ecosystem health
Dispersion	0–1, High = variability Habitat variability & heterogeneity
Diatom Soil Activity Index	>0.7 = High activity; <0.3 = Low activity Soil activity & human impact
Descy Diatom Index (DDI)	>4 = Clean; 3 = Moderate; <2 = Polluted Saprobity & general water quality
Chilean Diatom Index (IDCH)	1–2 = Good; 3 = Moderate; 4–5 = Polluted Chilean rivers water quality
Brazilian Regional Index	Region-specific bands Region-specific water quality

Fig. 15.1 Key parameters used for interpreting the results of various calculated ecological indices, providing reference thresholds and contextual meaning to assess water quality, trophic status, pollution sensitivity, and overall ecosystem condition.

CHAPTER 17

THE ULTIMATE TOOLKIT FOR PLOTTING ANY INDEX

"In the hands of a careful scientist, a simple plot can sketch the soul of an ecosystem."

While standard bar and line charts are foundational for visualizing individual indices, complex ecological monitoring often requires more dynamic graphical representations to synthesize multiple data points simultaneously. To enhance analytical efficiency, researchers can utilize custom, reusable R functions that automate the generation of specialized, publication-ready plots across any computed dataset.

Two highly effective visualization techniques in this advanced toolkit are circular (polar) plots and multi-index temporal/spatial trend lines:

Circular (Polar) Plots: These visualizations wrap standard bar charts around a central axis, providing an intuitive, clock-like representation of an index. This format is highly effective for visualizing closed-loop data, such as seasonal sampling cycles (e.g., months of the year) or spatial distributions around a central point (e.g., radial sampling around a point-source discharge).

Multi-Index Trend Plots: Ecological status is rarely defined by a single metric. A multi-index line plot allows for the simultaneous comparison of several indices (e.g., comparing IPS, TDI, and SLA) across a continuous spatial or temporal gradient. This visualization instantly reveals whether different ecological metrics correlate, diverge, or respond uniformly to environmental pressures.

The following reusable functions provide a versatile toolkit for generating these advanced visual structures with minimal redundant coding.

FUNCTION: PLOT ONE CIRCULAR INDEX

This function isolates a specific index from the dataset and projects the computed values onto a polar coordinate system.

```
plot_circular_index <- function(df, index_name) {
  df_index <- df %>% select(Sample, !!sym(index_name))

  ggplot(df_index, aes(x = Sample, y = .data[[index_name]])) +
    geom_bar(stat = "identity", fill = "#00BFC4") +
    coord_polar(start = 0) +
    theme_minimal() +
    labs(title = paste("Circular Plot -", index_name),
         y = "Score", x = "") +
    theme(axis.text.x = element_text(size = 10))
}
```

Run like this:

```
plot_circular_index(data, "IPS")
```

LINE PLOT OF MULTI-INDICES ACROSS SAMPLES

This function automatically pivots a dataset containing multiple indices into a long format, allowing ggplot2 to map each index as a distinct, color-coded trend line across sampling sites.

```
plot_index_line <- function(df) {
  df_long <- df %>%
    pivot_longer(cols = -c(X, Sample),
                 names_to = "Index",
                 values_to = "Score")

  ggplot(df_long, aes(x = Sample, y = Score, color = Index, group = Index)) +
```

```
geom_line(size = 1) +  
geom_point(size = 2) +  
theme_minimal() +  
labs(title = "Ecological Index Trends per Sample",  
      y = "Index Value") +  
theme(axis.text.x = element_text(angle = 45, hjust = 1))  
}
```

Run like this:

```
plot_index_line(data)
```

CHAPTER 18

DIATOM INDICES WITH DIATHOR – CODE, PLOTS, AND ECOLOGICAL INTERPRETATION

*"In every curve, every index, and every plot lies the truth of the waters
– silent, ancient, and waiting to be understood."*

Regional Considerations for Indian Waters

While indices like IPS, TDI, and IDCH are globally recognized, it is critical to acknowledge regional biases. Many of these indices were calibrated using European or South American river systems. When applying these metrics to Indian waters, researchers should cross-reference findings with local ecological knowledge and taxonomic records, as some endemic Indian taxa may lack established sensitivity scores in the default reference databases.

Feasibility of Global Diatom Indices in Indian Waters

Diatom-based ecological indices provided by DiaThor—such as IPS, EDI, ILM, and IDCH—were predominantly developed and calibrated for European or South American river systems. Their direct application to Indian waters requires a nuanced understanding of their feasibility:

IPS (Specific Pollution Sensitivity Index): This is one of the most widely applicable indices in India. Because it relies on a massive database of species, many of which are cosmopolitan (e.g., *Achnanthes minutissimum*, *Nitzschia palea*), it reliably detects organic pollution and general degradation in major Indian river systems. However, certain endemic tropical taxa may occasionally be flagged as "Excluded Taxa."

EDI (Ecological Diatom Index) & ILM (Italian Diatom Index): Calibrated for Mediterranean and specific European water systems, these indices can be used in Indian waters but should be treated as supplementary. The monsoonal hydrology in India creates extreme seasonal shifts in nutrient dilution and concentration, which these indices may sometimes misinterpret as abrupt ecological degradation rather than natural seasonal cycles.

IDCH (Chilean Diatom Index): Developed specifically for South American rivers, the feasibility of IDCH in India is low. The Andean ecological baselines differ significantly from Himalayan or Peninsular Indian rivers. It should be used strictly for broad comparative studies rather than for regional regulatory compliance.

For maximum accuracy in Indian biomonitoring, it is recommended to use the IPS as the primary metric, while corroborating results with trait-based analyses (like Van Dam classifications) until dedicated Indian indices are fully integrated into computational tools.

ECOLOGICAL INTERPRETATION OF THE KEY INDICES WITH CODE FOR GRAPH VISUALISATION

1. IPS (Pollution Sensitivity Index)

Definition:

The IPS evaluates the sensitivity of diatom species to organic pollution and general anthropogenic disturbances. Scores typically range from **1 (highly polluted)** to **20 (very clean)**. IPS is calibrated mainly for European rivers. Use with caution outside that context and always cross-check with local taxa response.

Interpretation:

- **High values (≥ 15):** Good ecological status, low organic pollution.
- **Low values (≤ 9):** Poor ecological conditions, heavy pollution.
- Often used in European water quality standards.

IPS Index

```
results_ips <- diat_ips(df)

ggplot(results_ips, aes(x = rownames(results_ips), y = IPS)) +
  geom_col(fill = "steelblue") +
  geom_hline(yintercept = 9, linetype = "dashed", color = "red") +
  geom_hline(yintercept = 15, linetype = "dashed", color = "green") +
  labs(title = "IPS Index by Site",
       x = "Sampling Site",
       y = "IPS Score (1-20)") +
  theme_minimal(base_size = 14)
```

Interpretation:

Sites above the green line are **ecologically healthy**, while those below the red line show **organic pollution stress**.

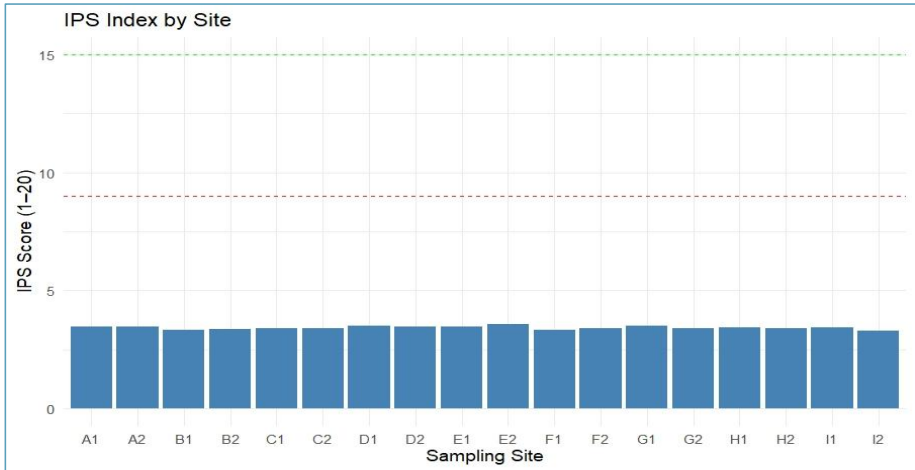


Fig. 17.1 Ecological interpretation of the *IPS (Specific Pollution Sensitivity Index)* by site, showing consistently low IPS scores (1–20 scale) across all sampling locations (A1–I2), indicative of **poor water quality and high pollution pressure**. Dashed reference lines represent standard IPS threshold values used for ecological status classification, highlighting that all sites fall well below acceptable ecological conditions.

2. TDI – Trophic Diatom Index

Definition:

TDI measures nutrient enrichment (eutrophication), particularly phosphorus. It ranges from 0 (nutrient-poor, oligotrophic) to 100 (nutrient-rich, eutrophic). TDI is highly sensitive to nutrient input and may overestimate trophic status in systems with naturally high productivity.

Interpretation:

Low values (<30): Clean, oligotrophic systems.

Moderate (30–50): Mesotrophic waters.

High (>50): Eutrophic, nutrient-enriched waters.

PLOTTING TDI (20-POINT SCALE)

```
# Calculate TDI index
results_tdi <- diat_tdi(df)

# Plot TDI Index with value labels (using TDI20)
ggplot(results_tdi, aes(x = rownames(results_tdi), y = TDI20)) +
  geom_col(fill = "darkorange") +
  geom_text(aes(label = round(TDI20, 1)),
    vjust = -0.5, size = 5, fontface = "bold") +
  geom_hline(yintercept = 30, linetype = "dashed", color = "green") + # Oligotrophic
  geom_hline(yintercept = 50, linetype = "dashed", color = "yellow") + # Mesotrophic
  geom_hline(yintercept = 70, linetype = "dashed", color = "red") +   # Eutrophic
  labs(title = "TDI (Trophic Diatom Index) by Site",
    x = "Sampling Site",
    y = "TDI Score (0–20)") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

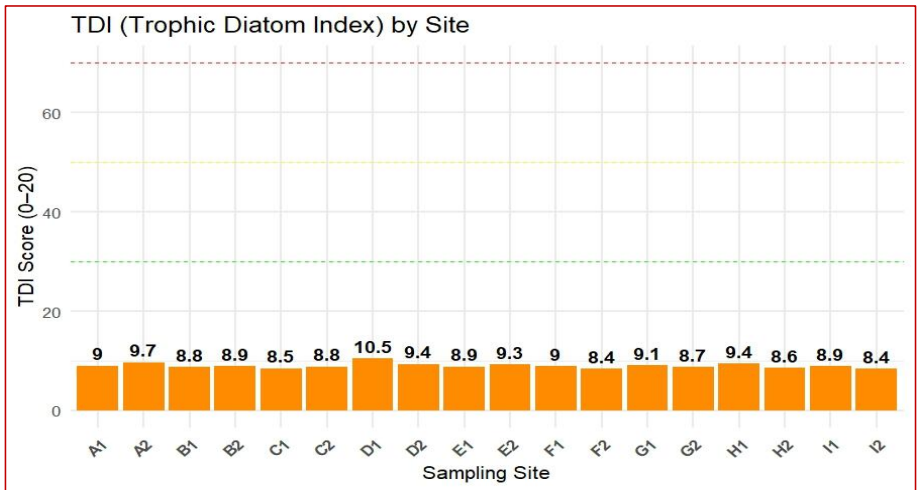



Fig. 17.2 Ecological interpretation of the TDI (Trophic Diatom Index) by site on a 20-point scale, illustrating site-wise variation in nutrient enrichment and trophic status. Reference threshold lines indicate standard TDI classification limits, enabling identification of sites reflecting eutrophic to hypertrophic conditions associated with elevated nutrient pressure.

PLOTTING TDI (100-POINT SCALE)

```
# Calculate TDI index
results_tdi <- diat_tdi(df)

# Plot TDI Index with value labels (using TDI100)
ggplot(results_tdi, aes(x = rownames(results_tdi), y = TDI100)) +
  geom_col(fill = "darkorange") +
  geom_text(aes(label = round(TDI100, 1)),
    vjust = -0.5, size = 5, fontface = "bold") +
  geom_hline(yintercept = 30, linetype = "dashed", color = "green") + # Oligotrophic
  geom_hline(yintercept = 50, linetype = "dashed", color = "yellow") + # Mesotrophic
  geom_hline(yintercept = 70, linetype = "dashed", color = "red") + # Eutrophic
  labs(title = "TDI (Trophic Diatom Index) by Site",
    x = "Sampling Site",
    y = "TDI Score (0-100)") +
  theme_minimal(base_size = 16) +
```

```
theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

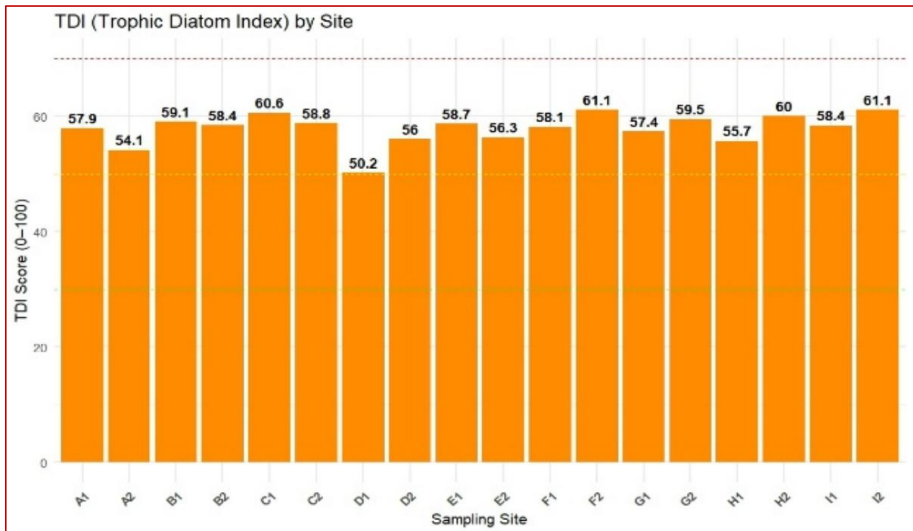


Fig. 17.3 Ecological interpretation of the TDI (Trophic Diatom Index) by site on a 100-point scale, illustrating site-wise variation in nutrient enrichment and trophic status. Reference threshold lines indicate standard TDI classification limits, enabling identification of sites reflecting eutrophic to hypertrophic conditions associated with elevated nutrient pressure.

Interpretation:

TDI20 is just a compressed version of **TDI100** (basically rescaled).

Use **TDI100** for clearer ecological interpretation, since it maps directly to trophic states:

<30 → oligotrophic (clean water)

30–50 → mesotrophic (moderate)

50 → eutrophic (nutrient-rich, impacted).

3. SLA – Sládeček Index

Definition:

Designed to evaluate organic pollution levels in rivers. Ranges from 1 (clean) to 5 (polluted).

Interpretation:

Low SLA (1–2): Good water quality.

High SLA (4–5): Heavily impacted by organic matter.

```
results_sla <- diat_sla(df)
```

```
ggplot(results_sla, aes(x = rownames(results_sla), y = SLA)) +
  geom_col(fill = "purple") +
  geom_hline(yintercept = 2, linetype = "dashed", color = "green") +
  geom_hline(yintercept = 4, linetype = "dashed", color = "red") +
  labs(title = "Sládeček Index by Site",
       x = "Sampling Site",
       y = "SLA Score (1-5)") +
  theme_minimal(base_size = 14)
```

Interpretation:

Higher SLA values = **organic pollution pressure**.

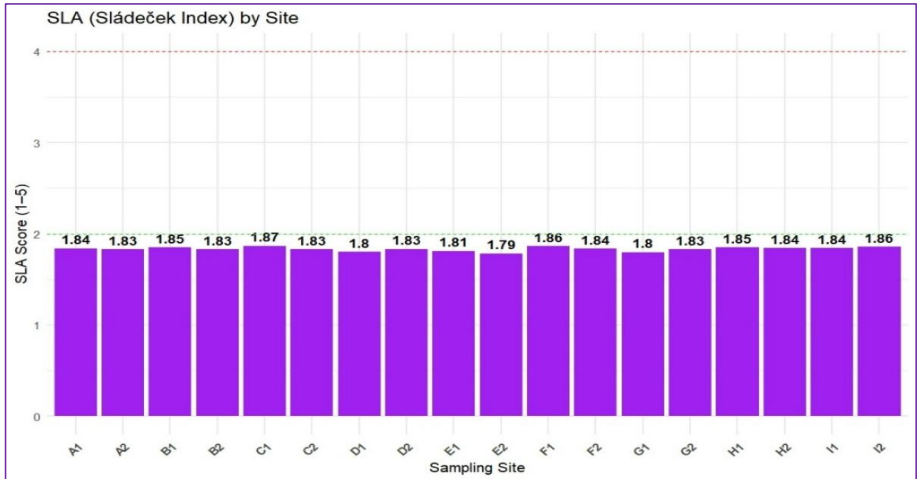


Fig. 17.4 Ecological interpretation of the SLA (Sládeček Index) by site, reflecting gradients of organic pollution and saprobity across sampling locations. Site-wise variations in SLA values indicate differences in ecological condition, with higher values corresponding to increased organic load and reduced water quality.

4. DDI – Descy Diatom Index

Definition: Another organic pollution index, operating on a 0–5 scale, often used in national biomonitoring. Originally designed for central European rivers. Local calibration is limited. Underrepresents rare species; focus is on dominant taxa.

Interpretation: A robust metric for pollution, especially organic enrichment.

- **High DDI (>4):** Clean systems.
- **Low DDI (<2):** Organic-rich or polluted waters.

```
# Calculate DDI index
results_ddi <- diat_ddi(df)

# Plot Descy Diatom Index with value labels
ggplot(results_ddi, aes(x = rownames(results_ddi), y = DDI)) +
  geom_col(fill = "seagreen") +
  geom_text(aes(label = round(DDI, 2)),
    vjust = -0.5, size = 5, fontface = "bold") +
  geom_hline(yintercept = 2, linetype = "dashed", color = "red") +
  geom_hline(yintercept = 4, linetype = "dashed", color = "green") +
  labs(title = "Descy Diatom Index by Site",
    x = "Sampling Site",
    y = "DDI Score (0-5)") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

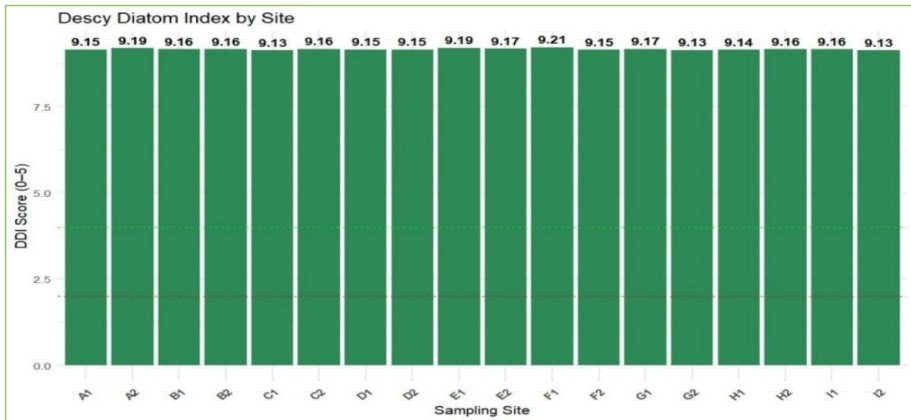


Fig. 17.5 Ecological interpretation of the DDI (Diatom Diversity Index) by site, illustrating variation in diatom community diversity across sampling locations. Higher DDI values indicate greater ecological stability and better water quality, whereas lower values reflect environmental stress and anthropogenic disturbance.

5. LOBO Index

Definition: Developed for tropical ecosystems, LOBO simplifies water quality classification. Range 1–5. Best applied within South American calibration zones.

Interpretation: Sensitive to multiple stressors, including organic matter.

1: Clean, pristine rivers.

5: Polluted, degraded waters.

```
results_lobo <- diat_lobo(df)
```

```
ggplot(results_lobo, aes(x = rownames(results_lobo), y = LOBO)) +
  geom_col(fill = "steelblue") +
  geom_text(aes(label = round(LOBO, 2)), vjust = -0.5, size = 5, fontface = "bold") +
  geom_hline(yintercept = 3, linetype = "dashed", color = "red") +
  labs(title = "LOBO Index by Site",
       x = "Sampling Site", y = "LOBO (1-5)") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

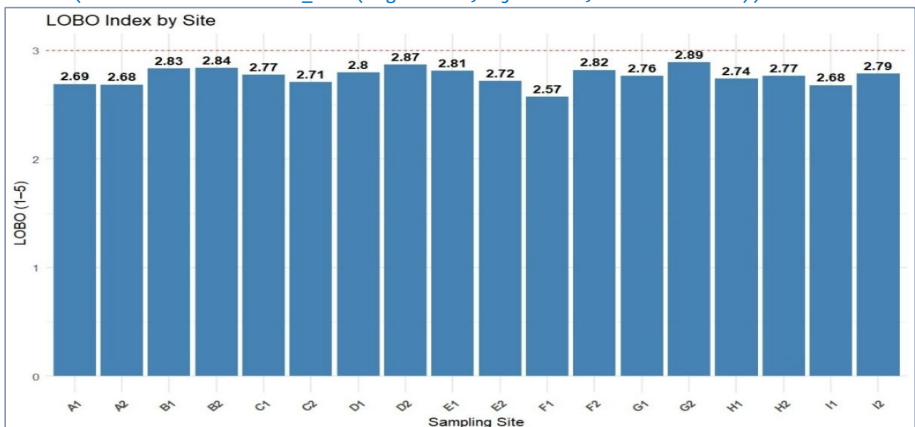


Fig. 17.6 Ecological interpretation of the LOBO Index by site, assessing organic pollution and overall water quality based on diatom community composition. Variations in LOBO values reflect differing degrees of environmental stress and anthropogenic impact across sampling locations.

6. EDI – Ecological Diatom Index

Definition: Mediterranean-specific index (range 0–5) focusing on ecological status under regional conditions. Not widely validated in tropical or subtropical waters.

Interpretation: Captures ecological status under European WFD (Water Framework Directive).

- **High (>4):** Very good ecological quality.
- **Low (<2):** Poor to bad water quality.

```
results_edi <- diat_edi(df)

ggplot(results_edi, aes(x = rownames(results_edi), y = EDI)) +
  geom_col(fill = "darkorange") +
  geom_text(aes(label = round(EDI, 2)), vjust = -0.5, size = 5, fontface = "bold") +
  geom_hline(yintercept = 2, linetype = "dashed", color = "red") +
  geom_hline(yintercept = 4, linetype = "dashed", color = "green") +
  labs(title = "Ecological Diatom Index (EDI) by Site",
       x = "Sampling Site", y = "EDI (0-5)") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

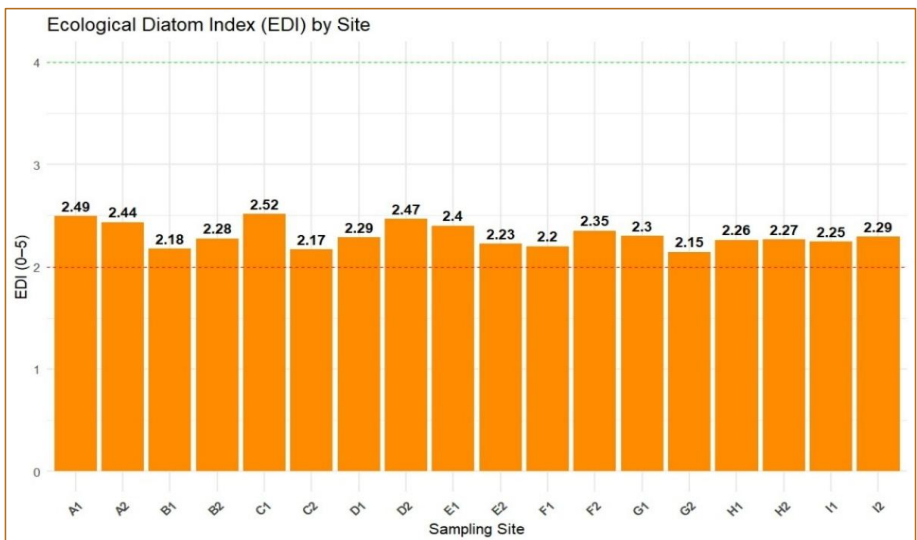


Fig. 17.7 Ecological interpretation of the EDI (Ecological Diatom Index) by site. Higher values indicate good ecological status and low pollution; lower values reflect degraded water quality and habitat disturbance.

7. ILM – Italian Diatom Index

Definition: Developed for Italian rivers; scores range from 1–18. Very regional; use outside Italy with care.

Interpretation: Strong indicator of general ecological status in Mediterranean rivers.

- **High (>12):** Healthy rivers.
- **Low (<6):** Polluted or impacted sites

```
results_ilm <- diat_ilm(df)

ggplot(results_ilm, aes(x = rownames(results_ilm), y = ILM)) +
  geom_col(fill = "purple") +
  geom_text(aes(label = round(ILM, 2)), vjust = -0.5, size = 5, fontface = "bold") +
```

```
labs(title = "Italian Diatom Index (ILM) by Site",
     x = "Sampling Site", y = "ILM (1-18)") +
theme_minimal(base_size = 16) +
theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

Interpretation:

Sites above ~12 is of **good ecological quality**, while <6 indicates **stress/pollution**.

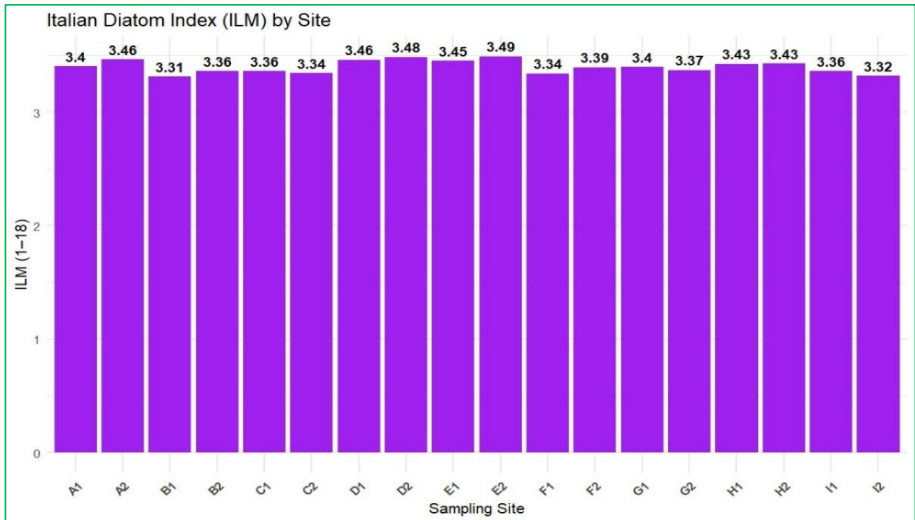


Fig. 17.8 Ecological interpretation of the ILM (Organic Pollution Index) by site. High values indicate minimal organic pollution and healthy water quality; low values reflect elevated organic loads and anthropogenic impact.

8. IDCH – Chilean Diatom Index

Definition: Region-specific for South American rivers. Don't over-interpret outside its calibration region.

- Range: 1–5
- 1 = clean, 5 = polluted

Interpretation: Specifically tailored for Chilean rivers, but useful in other South American systems.

```
results_idch <- diat_idch(df)
```

```
ggplot(results_idch, aes(x = rownames(results_idch), y = IDCH)) +
  geom_col(fill = "darkgreen") +
  geom_text(aes(label = round(IDCH, 2)), vjust = -0.5, size = 5, fontface = "bold") +
  labs(title = "Chilean Diatom Index (IDCH) by Site",
       x = "Sampling Site", y = "IDCH (1-5)") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

9. EPID – Epiphytic Index

Definition: Measures ecological status of **epiphytic diatom communities**. If sampling only sediments, this index may not apply.

- Scale: 1–5 (1 = clean, 5 = polluted)

Interpretation: Measures water quality as reflected by diatoms growing on plants (epiphytes).

```
results_epid <- diat_epid(df)
```

```
ggplot(results_epid, aes(x = rownames(results_epid), y = EPID)) +
  geom_col(fill = "tomato") +
  geom_text(aes(label = round(EPID, 2)), vjust = -0.5, size = 5, fontface = "bold") +
```

```
labs(title = "Epiphytic Diatom Index (EPID) by Site",  
      x = "Sampling Site", y = "EPID (1-5)") +  
theme_minimal(base_size = 16) +  
theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

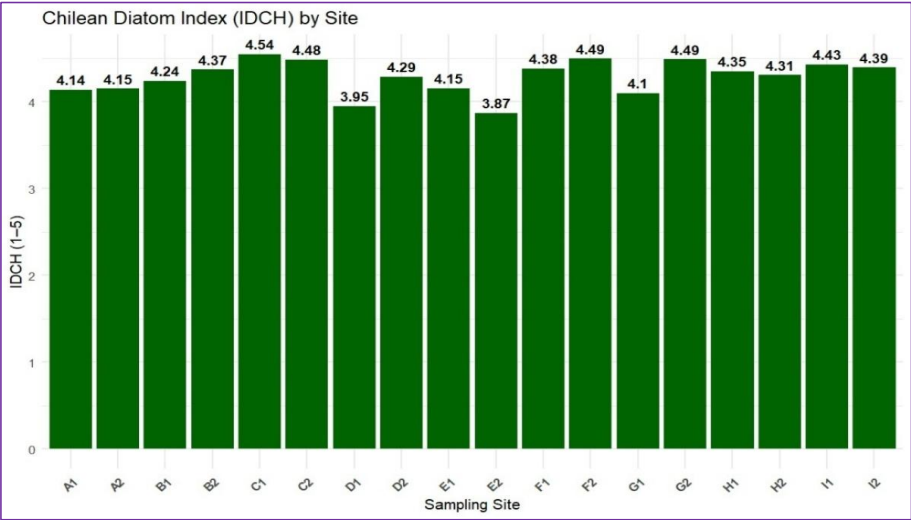


Fig. 17.9 Ecological interpretation of the IDCH (Diatom Community Homogeneity Index) by site. High values reflect diverse, balanced diatom communities and intact habitats; low values indicate dominance by tolerant species and habitat degradation

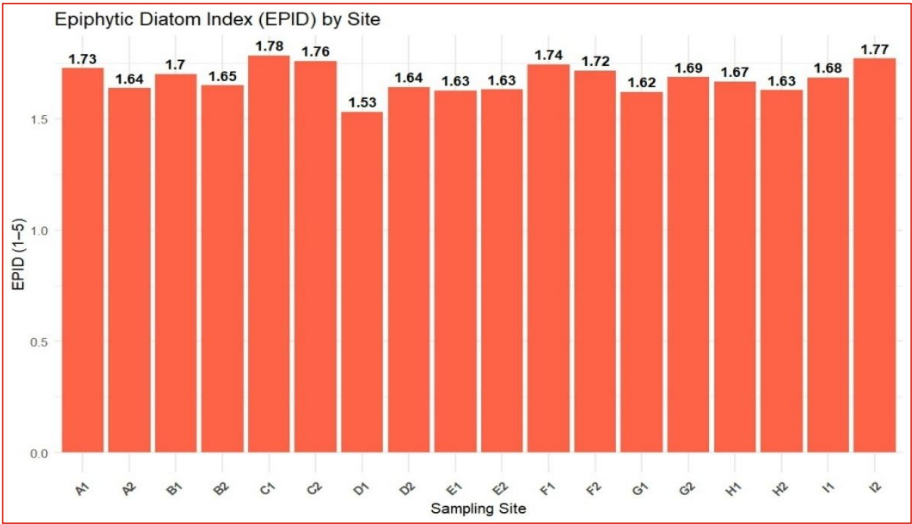


Fig. 17.10 Ecological interpretation of the EPID (Eutrophication Pressure Index) by site. High values indicate low nutrient pressure and oligotrophic conditions; low values reflect nutrient enrichment and eutrophication stress

10. IDP – Pedological Diatom Index

Definition: Evaluates **soil disturbance**. Sensitive to soil moisture variation.

- Higher values = more stable, undisturbed soils

Interpretation: Extends diatom biomonitoring to soils — helpful in agriculture, erosion, or land-use studies.


```
results_idp <- diat_idp(df)

ggplot(results_idp, aes(x = rownames(results_idp), y = IDP)) +
  geom_col(fill = "brown") +
  geom_text(aes(label = round(IDP, 2)), vjust = -0.5, size = 5, fontface = "bold") +
  labs(title = "Pedological Diatom Index (IDP) by Site",
       x = "Sampling Site", y = "IDP") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

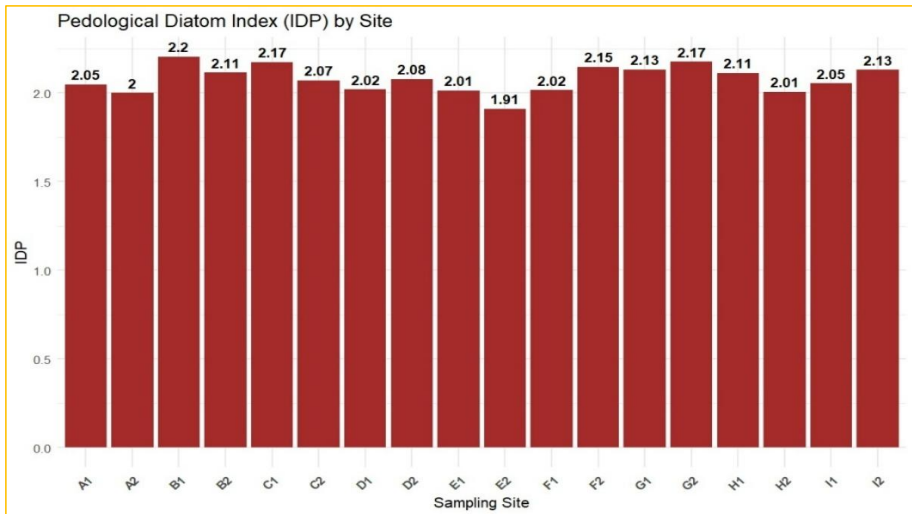


Fig. 17.11 Ecological interpretation of the IDP (Pollution Sensitivity Index) by site. High values show prevalence of pollution-sensitive diatom species, indicating low anthropogenic stress; low values reflect dominance by tolerant species under high pollution.

11. IDAP – Diatom Soil Activity Index

Definition: Assesses **biological activity in soils**. Best used in comparative studies (before vs. after land disturbance).

- Range: 0–1
- 0.7 = biologically active
- <0.3 = poor

Interpretation: Indicates microbial and diatom productivity in soils.

```
results_idap <- diat_idap(df)

ggplot(results_idap, aes(x = rownames(results_idap), y = IDAP)) +
  geom_col(fill = "darkred") +
  geom_text(aes(label = round(IDAP, 2)), vjust = -0.5, size = 5, fontface = "bold") +
  labs(title = "Diatom Soil Activity Index (IDAP) by Site",
       x = "Sampling Site", y = "IDAP (0-1)") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

12. SPEAR – Herbicide Sensitivity Index

Definition:

Identifies impact of **toxic contamination** (herbicides). Requires careful species identification; sensitive to misclassifications.

- Higher = more sensitive taxa present → less pesticide impact

Interpretation: Targets pesticide/herbicide contamination specifically.

```
results_spear <- diat_spear(df)

ggplot(results_spear, aes(x = rownames(results_spear), y = SPEAR)) +
  geom_col(fill = "gold") +
  geom_text(aes(label = round(SPEAR, 2)), vjust = -0.5, size = 5, fontface = "bold") +
  labs(title = "SPEAR Index by Site",
       x = "Sampling Site", y = "SPEAR") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

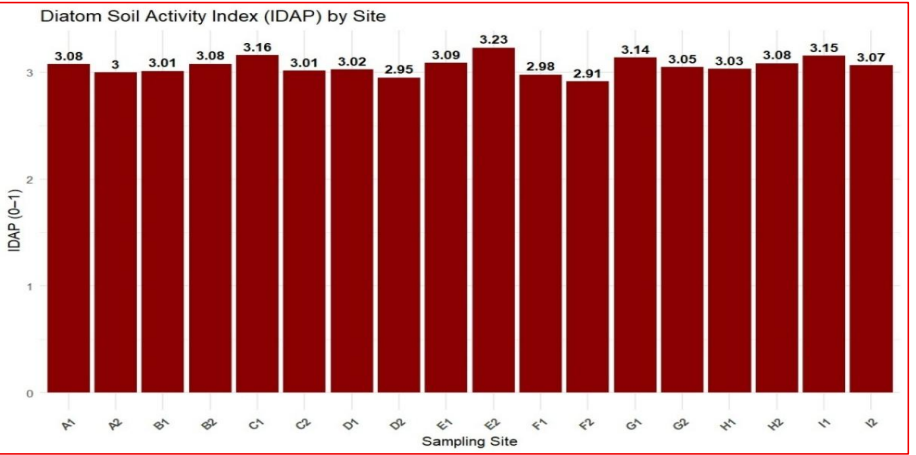


Fig. 17.12 Ecological interpretation of the IDAP (Anthropogenic Pressure Index) by site. High values indicate minimal human-induced stress and stable habitats; low values reflect significant anthropogenic pressure and habitat alteration.

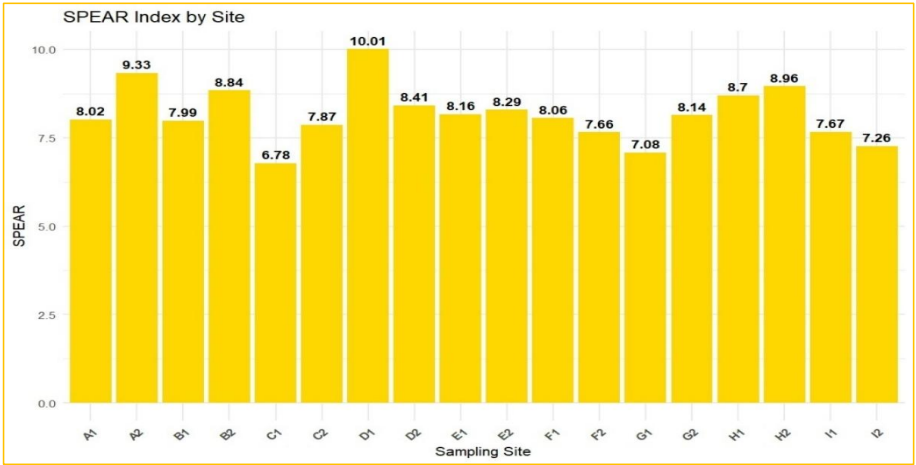


Fig. 17.13 Ecological interpretation of the SPEAR (Species at Risk—Pesticide Sensitivity) by site. High values indicate healthy, pesticide-free ecosystems; low values reflect pesticide contamination and reduced ecological integrity.

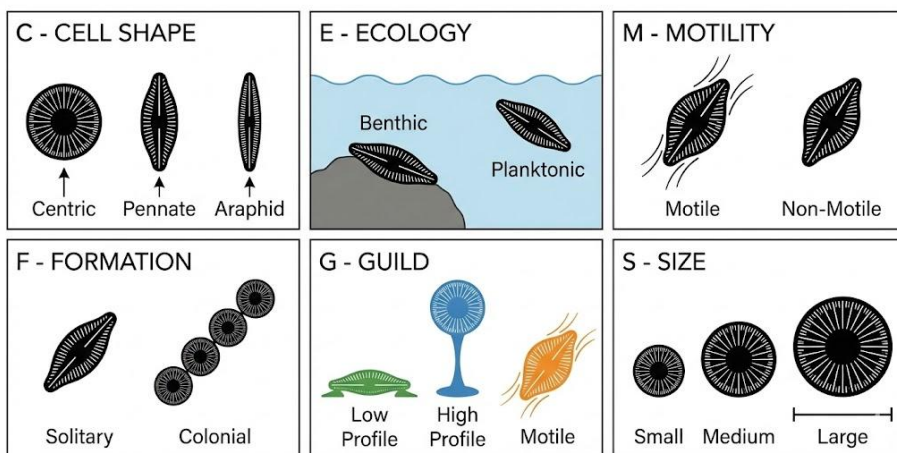
13. CEMFGS (Cell Shape / Ecology / Motility / Formation / Guild / Size)

Definition:

The CEMFGS classification is a comprehensive system that categorizes diatoms into functional groups based on a combination of traits:

- **C** – Cell Shape (morphological form of the frustule)
- **E** – Ecology (preferred habitat or tolerance range)
- **M** – Motility (ability to move across substrates)
- **F** – Formation (growth form, such as colonial or solitary)
- **G** – Guild (functional groups like motile, high-profile, low-profile, planktonic)
- **S** – Size class (relative cell size, influencing ecological strategy)

THE CEMFGS FRAMEWORK: DIATOM FUNCTIONAL TRAITS



This approach integrates multiple biological characteristics, providing a more **holistic functional perspective** of diatom assemblages. It is particularly useful for understanding how communities adapt to **disturbances, nutrient enrichment, and habitat heterogeneity**. For instance, motile guilds may dominate in sediment-rich environments, while large-sized high-profile forms may be abundant in stable, nutrient-poor habitats. Interpretation is descriptive, not “good/bad” like indices. Needs cross-checking with other metrics.

Interpretation:

- **Higher proportion of motile forms (M)** → often linked to disturbed or polluted habitats.
- **Balanced LS (low-profile) and HS (high-profile) groups** → suggest ecological stability.
- **Size distribution shifts** → can indicate stress, with smaller forms prevailing in degraded systems.
- **Indeterminate (Indet) fraction** → represents taxa not fitting into clear categories, often highlighting gaps in ecological knowledge.

#CODE:

```
results_cemfgs_rb <- diat_cemfgs_rb(df)
```

```
ggplot(results_cemfgs_rb, aes(x = rownames(results_cemfgs_rb), y = CEMFGS)) +
```

```
geom_col(fill = "skyblue") +
geom_text(aes(label = round(CEMFGS, 2)), vjust = -0.5, size = 5, fontface = "bold") +
labs(title = "CEMFGS Index by Site",
      x = "Sampling Site", y = "CEMFGS") +
theme_minimal(base_size = 16) +
theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

TROUBLESHOOTING ERRORS:

The error is happening because in your `results_cemfgs_rb` object, the column is probably not named `CEMFGS`, but something else. That's why `geom_col()` can't find `CEMFGS`.

Here's how you can fix it step by step:

Step 1: Inspect the results

```
head(results_cemfgs_rb)
colnames(results_cemfgs_rb)
```

- This will show you the actual column names. Once you know the correct column, you can plug it into the plot.
- The CEMFGS index values are stored in a column with the unusual name "CEMFGS class Indet" (with spaces).
 - The taxa count used is in "CEMFGS Taxa used".
- Since R doesn't like spaces in column names when directly referenced, you need to wrap them in backticks (`).

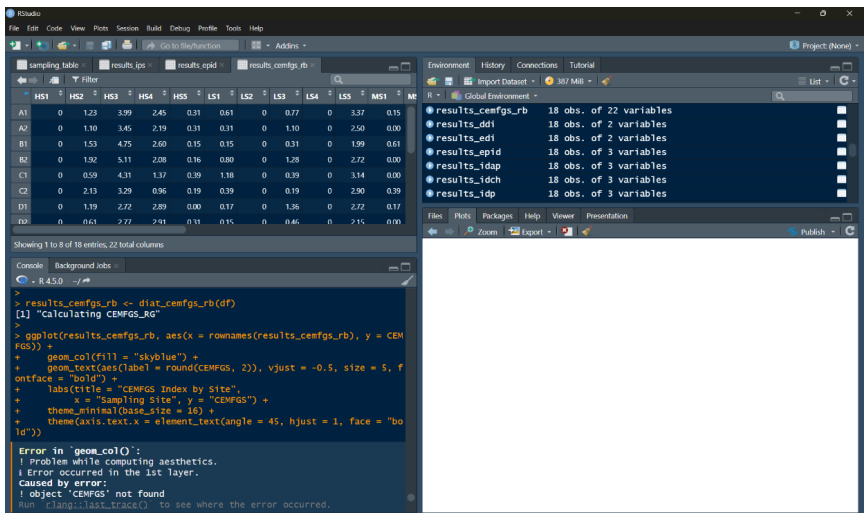


Fig. 17.14 RStudio interface illustrating the CEMFGS framework

```
#Plotting CEMFGS index scores
library(ggplot2)
```

```
ggplot(results_cemfgs_rb, aes(x = rownames(results_cemfgs_rb), y = `CEMFGS class Indet`))
+
geom_col(fill = "steelblue") +
geom_text(aes(label = round(`CEMFGS class Indet`, 2)),
          vjust = -0.5, size = 5, fontface = "bold") +
labs(title = "CEMFGS Index by Site",
      x = "Sampling Site",
      y = "CEMFGS Score") +
theme_minimal(base_size = 16) +
```

```
theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

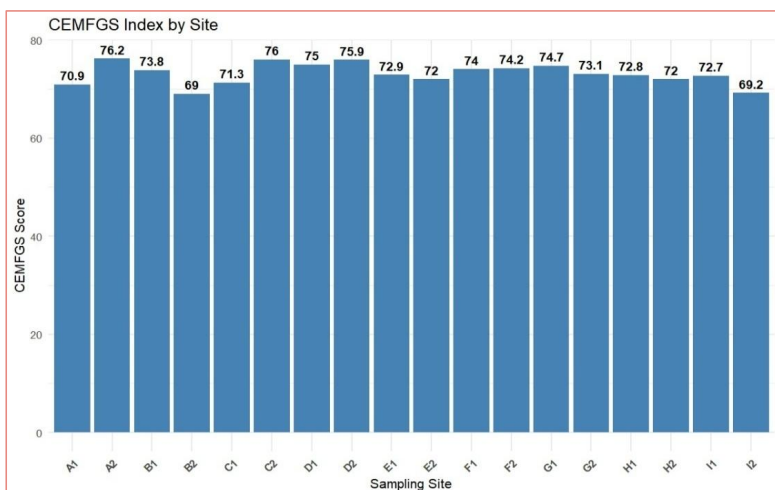


Fig. 17.15 Plot showing CEMFGS index scores across sampling sites,

```
#Plotting number of taxa used
ggplot(results_cemfgs_rb, aes(x = rownames(results_cemfgs_rb), y = `CEMFGS Taxa used`))
+
  geom_col(fill = "darkorange") +
  geom_text(aes(label = `CEMFGS Taxa used`),
            vjust = -0.5, size = 5, fontface = "bold") +
  labs(title = "Taxa Used in CEMFGS Index Calculation",
       x = "Sampling Site",
       y = "Number of Taxa") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"))
```

#Since our results_cemfgs_rb has columns for each guild type (HS, LS, MS, PS), we can reshape it to a long format and then plot a stacked bar chart.

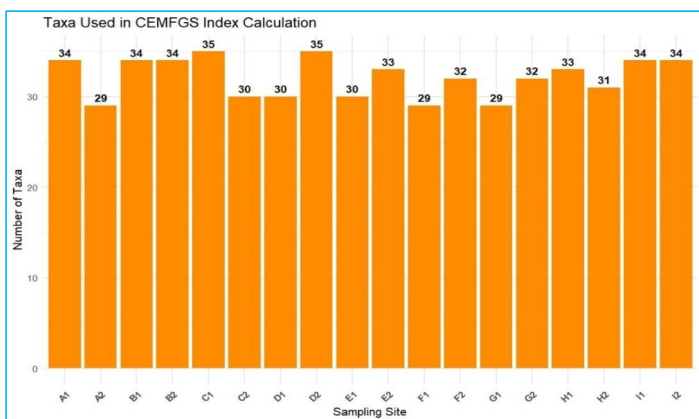


Fig. 17.16 Plot showing the number of diatom taxa used for index calculation across sampling sites

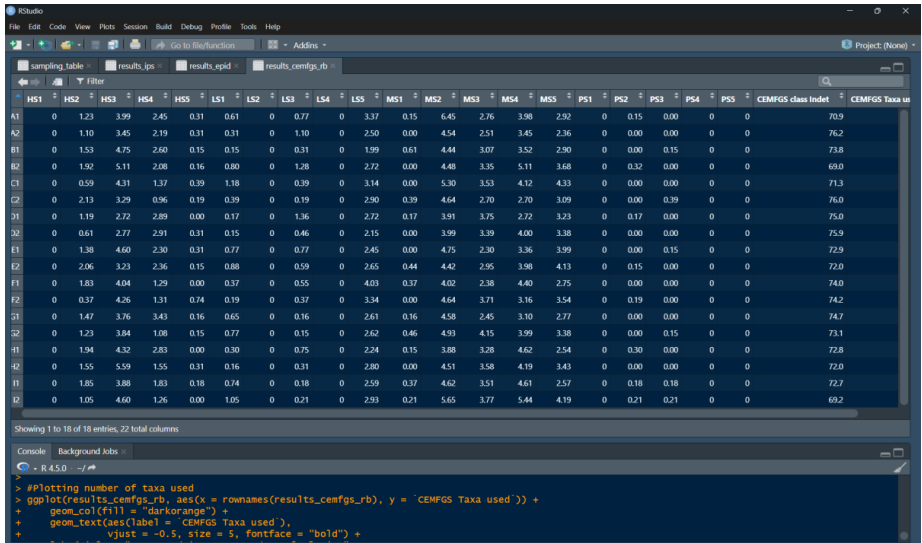


Fig. 17.17 RStudio working interface displaying the relative contribution of CEMFGS guild types—High-profile (HS), Low-profile (LS), Motile (MS), and Planktonic (PS)—across sampling sites. The plot illustrates site-wise shifts in diatom ecological guild structure, reflecting differences in habitat stability, nutrient conditions, and anthropogenic disturbance.

```

library(ggplot2)
library(dplyr)
library(tidyr)

# Reshape to long format (keep only HS, LS, MS, PS groups)
results_cemfgs_long <- results_cemfgs_rb %>%
  tibble::rownames_to_column("Site") %>%
  select(Site, HS1:PS5) %>% # select only guild columns
  pivot_longer(~Site, names_to = "Guild", values_to = "Value")

# Plot stacked bar chart
ggplot(results_cemfgs_long, aes(x = Site, y = Value, fill = Guild)) +
  geom_bar(stat = "identity") +
  labs(title = "CEMFGS Functional Guild Composition by Site",
       x = "Sampling Site",
       y = "Relative Abundance (%)",
       fill = "Guild Class") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
        plot.title = element_text(face = "bold", size = 18))

```

What this does:

Each site is on the x-axis.

The y-axis shows percentage contribution.

Each color bar shows the proportion of HS, LS, MS, PS categories.

If you'd prefer dodged bars instead of stacked (so you can compare side by side), just replace:

```
geom_bar(stat = "identity", position = "dodge")
```

Since the stacked bars have many segments, the labels inside each segment get cramped and unreadable. Instead, let's try two clearer options:

Option 1: Show total values above bars (cleaner view)**CODE:**

```
library(ggplot2)
library(dplyr)
library(tidyr)

results_cemfgs_long <- results_cemfgs_rb %>%
  tibble::rownames_to_column("Site") %>%
  select(Site, HS1:PS5) %>%
  pivot_longer(~Site, names_to = "Guild", values_to = "Value")

# Summarize totals per site
totals <- results_cemfgs_long %>%
  group_by(Site) %>%
  summarise(Total = sum(Value, na.rm = TRUE))

ggplot(results_cemfgs_long, aes(x = Site, y = Value, fill = Guild)) +
  geom_bar(stat = "identity") +
  geom_text(data = totals, aes(x = Site, y = Total + 2, label = round(Total, 1)),
            inherit.aes = FALSE, size = 4, fontface = "bold") +
  labs(title = "CEMFGS Functional Guild Composition by Site",
       x = "Sampling Site", y = "Relative Abundance (%)",
       fill = "Guild Class") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
        plot.title = element_text(face = "bold", size = 18))
```

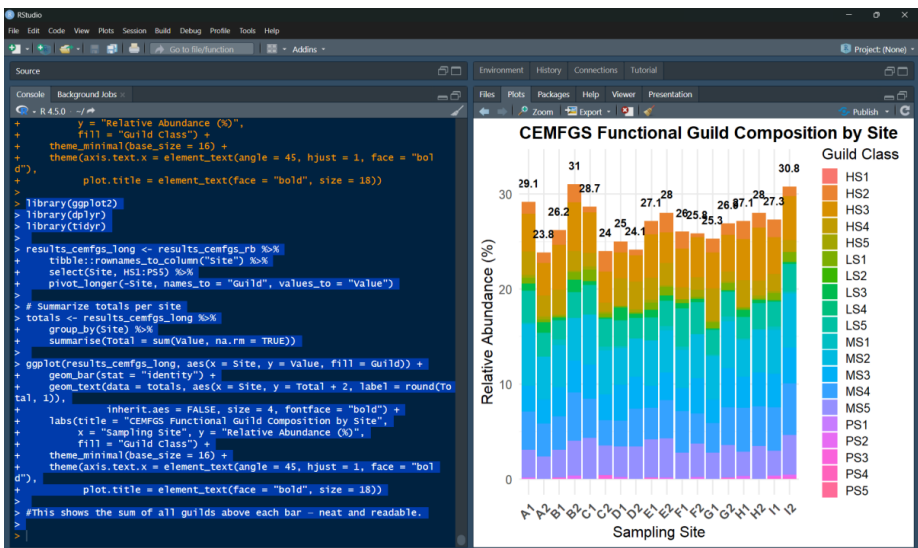


Fig. 17.18 Visualization of CEMFGS ecological guild composition (HS, LS, MS, PS) across sampling sites generated after reshaping guild data into long format, highlighting functional shifts in diatom communities along environmental gradients.

#This shows the sum of all guilds above each bar – neat and readable.

Option 2: Switch to a dodged barplot (side-by-side bars with values above each bar)**#CODE:**


```
ggplot(results_cemfgs_long, aes(x = Site, y = Value, fill = Guild)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9)) +
  geom_text(aes(label = round(Value,1)),
            position = position_dodge(width = 0.9),
            vjust = -0.3, size = 3) +
  labs(title = "CEMFGS Functional Guild Composition by Site",
        x = "Sampling Site", y = "Relative Abundance (%)",
        fill = "Guild Class") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
        plot.title = element_text(face = "bold", size = 18))
```

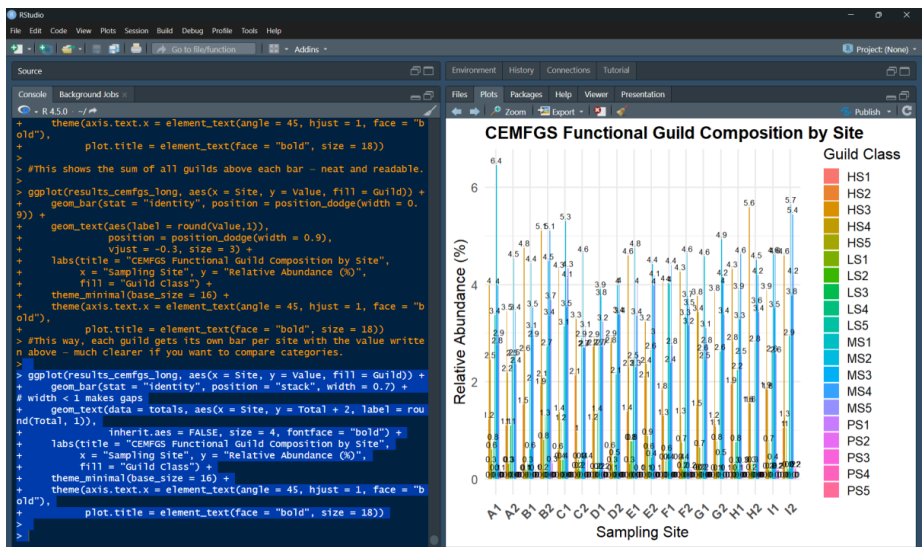


Fig. 17.19 Dodged bar plot showing site-wise distribution of CEMFGS ecological guild types—High-profile (HS), Low-profile (LS), Motile (MS), and Planktonic (PS). Side-by-side bars allow direct comparison of guild-wise contributions across sampling sites, highlighting functional shifts in diatom community structure in response to environmental conditions.

#This way, each guild gets its own bar per site with the value written above – much clearer if you want to compare categories.

If you want the bars separated a bit instead of being glued together. That can be done by adjusting the **width** argument in **geom_bar()** or **geom_col()**. **Smaller width values = more space between bars.**

Here's how you can do it for both cases:

Option 1: Totals above grouped bars with gaps

```
ggplot(results_cemfgs_long, aes(x = Site, y = Value, fill = Guild)) +
  geom_bar(stat = "identity", position = "stack", width = 0.7) + # width < 1 makes gaps
  geom_text(data = totals, aes(x = Site, y = Total + 2, label = round(Total, 1)),
            inherit.aes = FALSE, size = 4, fontface = "bold") +
  labs(title = "CEMFGS Functional Guild Composition by Site",
        x = "Sampling Site", y = "Relative Abundance (%)",
        fill = "Guild Class") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
        plot.title = element_text(face = "bold", size = 18))
```

Option 2: Side-by-side bars (dodged) with gaps

```
ggplot(results_cemfgs_long, aes(x = Site, y = Value, fill = Guild)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.7) +
  geom_text(aes(label = round(Value,1)),
            position = position_dodge(width = 0.8),
            vjust = -0.3, size = 3) +
  labs(title = "CEMFGS Functional Guild Composition by Site",
        x = "Sampling Site", y = "Relative Abundance (%)",
        fill = "Guild Class") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
        plot.title = element_text(face = "bold", size = 18))
```

- If you want a **clearer separation between sites**, keep `width = 0.6 or even 0.5`.
- If you want **space between guilds inside each site**, adjust the `position_dodge(width = ...)` parameter.

13. Van Dam Classification**Definition:**

- The Van Dam classification groups diatoms by ecological preferences such as saprobity, trophic state, oxygen requirements, salinity, and nitrogen uptake. It is one of the most widely used ecological trait systems to interpret environmental gradients. Qualitative, but very insightful when paired with chemical data. Guilds are qualitative and best used as complementary to numeric indices.
- **Saprobity:** Clean → Moderately polluted → Heavily polluted
- **Oxygen:** Aerophilous → Facultative → Anoxic tolerant
- **Salinity:** Freshwater → Brackish → Saline
- **Trophic State:** Oligotrophic → Mesotrophic → Eutrophic

Interpretation: Provides trait-based ecological tendencies across gradients. Instead of a single score, guilds show *ecological tendencies* (e.g., aerophilous guild dominance suggests intermittent drying).

#CODE:

```
# Run Van Dam classification
results_vandam <- diat_vandam(df)

# Reshape into long format for plotting
results_vandam_long <- results_vandam %>%
  tidyr::pivot_longer(-Site, names_to = "Trait", values_to = "Value")

# Plot ecological preferences
ggplot(results_vandam_long, aes(x = Site, y = Value, fill = Trait)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Van Dam Ecological Preferences by Site",
        x = "Sampling Site",
        y = "Relative Abundance (%)",
        fill = "Ecological Trait") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
        plot.title = element_text(size = 18, face = "bold"))
```

```
Error in `tidyr::pivot_longer()` :
! Can't select columns that doesn't exist.
```

✗ Column `Site` doesn't exist.

That means the output of `diat_vandam(df)` doesn't contain a column named `Site`. Instead, your site names (A1, A2, B1 ...) are probably stored as rownames rather than in a dedicated column.

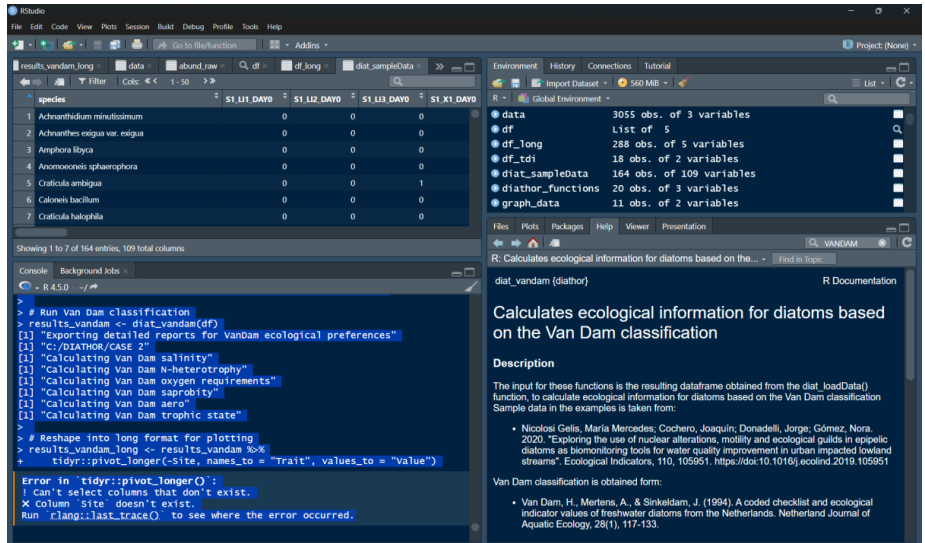


Fig. 17.20 RStudio working interface illustrating data reshaping using `pivot_longer()`

Quick Fix

```
# Before pivoting, convert rownames into a column:
library(dplyr)
library(tidyr)
library(ggplot2)

# Run Van Dam classification
results_vandam <- diat_vandam(df)

# Convert rownames into a "Site" column
results_vandam <- results_vandam %>%
  tibble::rownames_to_column(var = "Site")

# Reshape into long format for plotting
results_vandam_long <- results_vandam %>%
  pivot_longer(-Site, names_to = "Trait", values_to = "Value")

# Plot ecological preferences
ggplot(results_vandam_long, aes(x = Site, y = Value, fill = Trait)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  labs(title = "Van Dam Ecological Preferences by Site",
       x = "Sampling Site", y = "Relative Abundance (%)",
       fill = "Ecological Trait") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
```

```
plot.title = element_text(size = 18, face = "bold"))
```

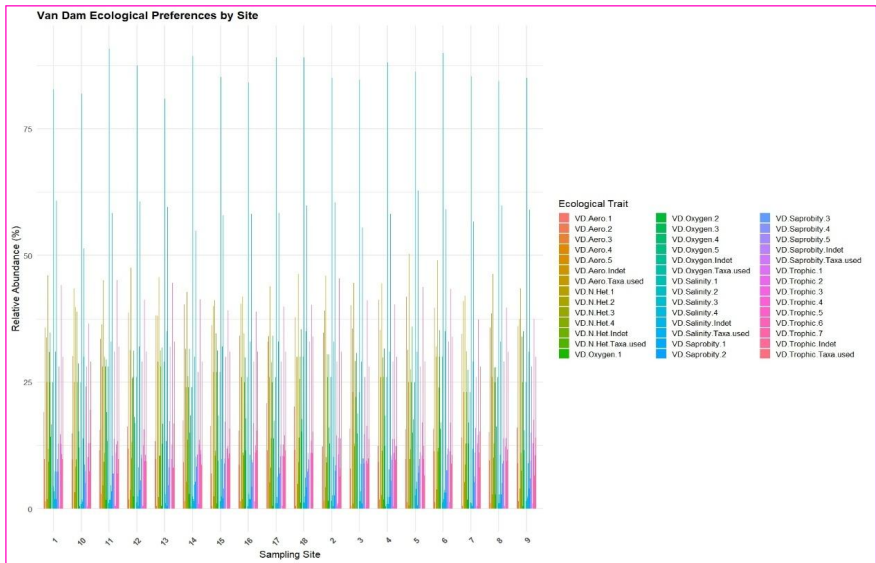


Fig. 17.21 Plot showing reshaped CEMFGS ecological guild data into long format for visualization.

Why this works:

`rownames_to_column()` moves the rownames (your site IDs like A1, A2, ...) into a real column called Site.

Now `pivot_longer(-Site, ...)` works, because it can keep that column fixed while melting the rest.

14. Ecological Guilds

Definition:

Diatoms are grouped into functional guilds:

- Motile (able to move along surfaces)
- High-profile (tall, competitive forms)
- Low-profile (small, resistant species)
- Planktonic (free-floating)

#CODE:

```
# Run Guild classification
results_guilds <- diat_guilds(df)

# Reshape for plotting
results_guilds_long <- results_guilds %>%
  tidyr::pivot_longer(-Site, names_to = "Guild", values_to = "Abundance")

# Plot Guild composition
ggplot(results_guilds_long, aes(x = Site, y = Abundance, fill = Guild)) +
  geom_bar(stat = "identity") +
  labs(title = "Ecological Guild Composition by Site",
       x = "Sampling Site",
       y = "Relative Abundance (%)",
       fill = "Guild Type") +
```

```
theme_minimal(base_size = 16) +
theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
      plot.title = element_text(size = 18, face = "bold"))
```

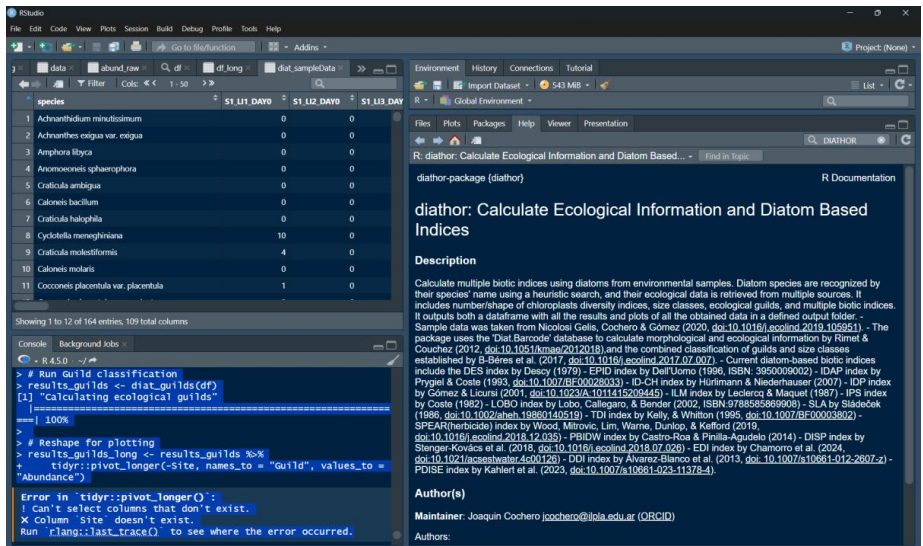


Fig. 17.22 RStudio working interface illustrating visualization of diatom guild composition derived from CEMFGS analysis.

The error is the same story as before: our `results_guilds` object doesn't have a column named `Site`. Instead, the sample IDs (A1, A2, etc.) are stored as rownames.

You just need to convert rownames into a proper column before pivoting. Here's the fixed code for the Guilds plot:

```
library(dplyr)
library(tidyr)
library(ggplot2)

# Run Guild classification
results_guilds <- diat_guilds(df)

# Convert rownames to a column named "Site"
results_guilds <- results_guilds %>%
  tibble::rownames_to_column(var = "Site")

# Reshape for plotting
results_guilds_long <- results_guilds %>%
  pivot_longer(~Site, names_to = "Guild", values_to = "Abundance")

# Plot Guild composition
ggplot(results_guilds_long, aes(x = Site, y = Abundance, fill = Guild)) +
  geom_bar(stat = "identity", width = 0.7) +
  labs(title = "Ecological Guild Composition by Site",
       x = "Sampling Site", y = "Relative Abundance (%)",
       fill = "Guild Type") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
        plot.title = element_text(size = 18, face = "bold"))
```

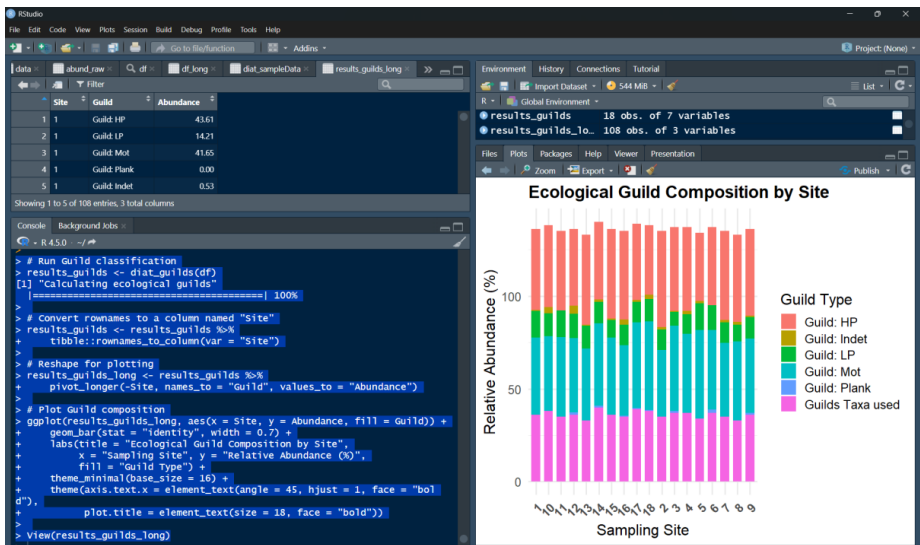


Fig. 17.23 Plot showing reshaped CEMFGS ecological guild composition data by site

- This ensures Site exists as a column, so `pivot_longer(~Site, ...)` will work. You'll get grouped bars (motive, low-profile, high-profile, planktonic) for each site.
- Wider bars (width = 0.7) make values clearer.

15. Size Classes

Definition:

This groups diatoms by their cell size (small → medium → large). Size structure often reflects ecological pressures such as nutrient status, grazing, and flow conditions.

```

library(dplyr)
library(tidyr)
library(ggplot2)
library(tibble) # for rownames_to_column()

```

```

# Run Size analysis
results_size <- diat_size(df)

```

```

# Move rownames into a column named "Site"
results_size <- results_size %>%
  rownames_to_column(var = "Site")

```

```

# Reshape for plotting
results_size_long <- results_size %>%
  pivot_longer(~Site, names_to = "SizeClass", values_to = "Proportion")

```

```

# Plot Size Class Distribution
ggplot(results_size_long, aes(x = Site, y = Proportion, fill = SizeClass)) +
  geom_bar(stat = "identity") +
  labs(title = "Diatom Size Class Distribution by Site",
       x = "Sampling Site",
       y = "Proportion (%)",
       fill = "Size Class") +

```

```
theme_minimal(base_size = 16) +
theme(
  axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
  plot.title = element_text(size = 18, face = "bold")
)
```

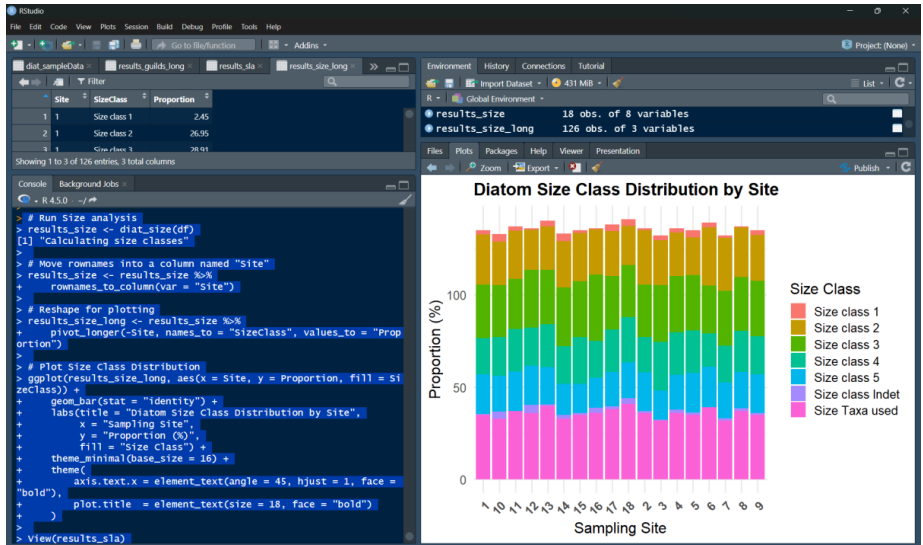


Fig. 17.24 RStudio interface showing diatom size-class composition across sampling sites.

16. Diversity Metrics

Definition:

Diatom diversity is measured with species richness, Shannon diversity index, and evenness. These reflect ecosystem health, resilience, and balance.

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(tibble) # for rownames_to_column()

# Run Diversity metrics
results_diversity <- diat_diversity(df)

# Move rownames into a "Site" column
results_diversity <- results_diversity %>%
  rownames_to_column(var = "Site")

# Reshape for plotting (everything except Site is gathered)
results_diversity_long <- results_diversity %>%
  pivot_longer(cols = -Site, names_to = "Metric", values_to = "Value")

# Plot Diversity Metrics
ggplot(results_diversity_long, aes(x = Site, y = Value, fill = Metric)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Diversity Metrics (Richness, Shannon, Evenness)",
       x = "Sampling Site",
```



```

y = "Value",
fill = "Metric") +
theme_minimal(base_size = 16) +
theme(
  axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
  plot.title = element_text(size = 18, face = "bold")
)

```

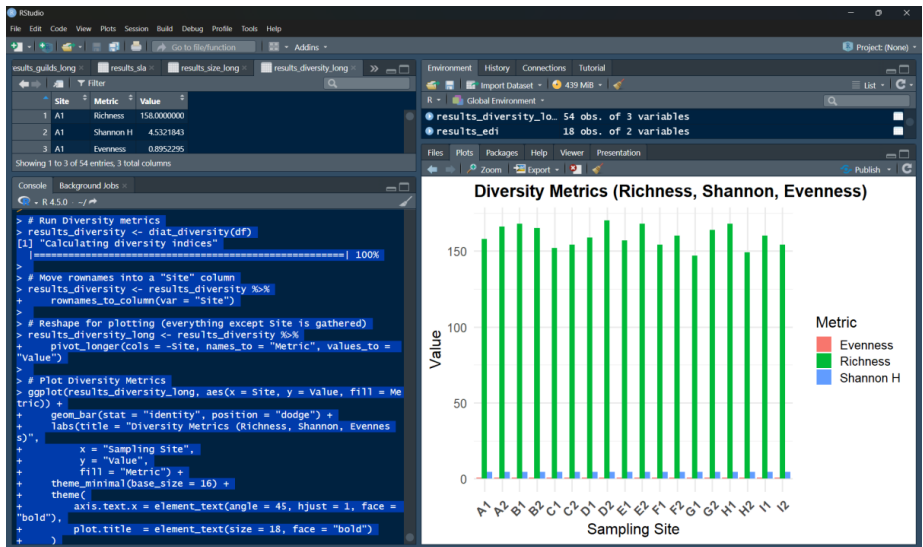


Fig. 17.25 Site-wise variation in diatom diversity metrics as visualized in the RStudio interface, reflecting differences in ecological condition and environmental stress.

17. Dispersion (DISP)

Definition:

Dispersion measures variance and heterogeneity among diatom communities across sites, indicating ecological stability or disturbance.

```

# Run Dispersion analysis
results_disp <- diat_disp(df)

# Plot Dispersion
ggplot(results_disp, aes(x = rownames(results_disp), y = Dispersion)) +
  geom_col(fill = "tomato") +
  geom_text(aes(label = round(Dispersion, 2)), vjust = -0.5, size = 5, fontface = "bold")
+
  labs(title = "Community Dispersion Across Sites",
        x = "Sampling Site",
        y = "Dispersion Value") +
  theme_minimal(base_size = 16) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
        plot.title = element_text(size = 18, face = "bold"))

```

Error message object 'Dispersion' not found means that the output of `diat_disp(df)` doesn't actually have a column named Dispersion. Instead, it likely has either a single unnamed column or something like "disp".

Let's fix it step by step:

1. Inspect column names first

```
colnames(results_disp)
head(results_disp)
```

This will show what the actual column is called.

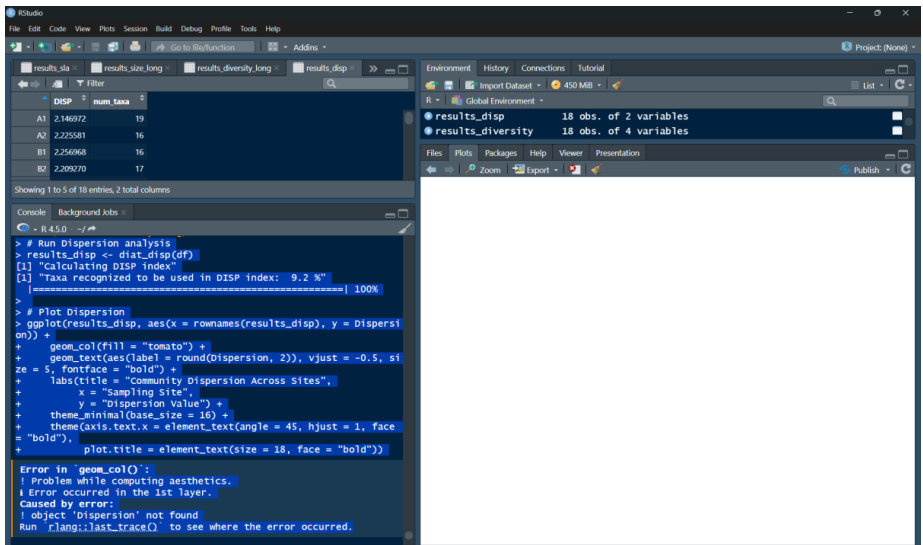


Fig. 17.26 RStudio interface illustrating calculation and visualization of Dispersion (DISP) metrics used to assess variability in diatom assemblages across sites.

2. Make sure Site exists

Since your sites (A1, A2, etc.) are stored as rownames, we need to move them into a column:

```
library(tibble)
results_disp <- results_disp %>%
  rownames_to_column(var = "Site")
```

3. Rename the dispersion column (if needed)

If your column is unnamed or has a weird name, rename it to "Dispersion":

```
colnames(results_disp)[2] <- "Dispersion"
`colnames(results_disp) head(results_disp)`
> colnames(results_disp) [1] "DISP" "num_taxa"
> head(results_disp) DISP num_taxa A1 2.146972 19 A2
2.225581 16 B1 2.256968 16 B2 2.209270 17 C1 2.157316 17
C2 2.212038 15
```

```
> colnames(results_disp)
[1] "DISP" "num_taxa"
> head(results_disp)
DISP num_taxa
A1 2.146972 19
A2 2.225581 16
B1 2.256968 16
B2 2.209270 17
C1 2.157316 17
C2 2.212038 15
```

Fig. 17.27 Screenshot of result displayed in RStudio interface.

So, the column you want to plot is called DISP (not Dispersion). The fix is simple – let's use that column directly.

4. Plot correctly

Here's the corrected plotting code:

```
library(tibble)
library(ggplot2)

# Move rownames to a "Site" column
results_disp <- results_disp %>%
  rownames_to_column(var = "Site")

# Plot DISP values by Site
ggplot(results_disp, aes(x = Site, y = DISP)) +
  geom_col(fill = "tomato") +
  geom_text(aes(label = round(DISP, 2)),
            vjust = -0.5, size = 5, fontface = "bold") +
  labs(title = "Community Dispersion Across Sites",
       x = "Sampling Site",
       y = "Dispersion Value (DISP)") +
  theme_minimal(base_size = 16) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
    plot.title = element_text(size = 18, face = "bold")
  )
```

What this does:

- `rownames_to_column()` ensures you have Site as a column.
- `geom_text()` adds the DISP value on top of each bar.
- Bigger, bold fonts make it publication-ready.

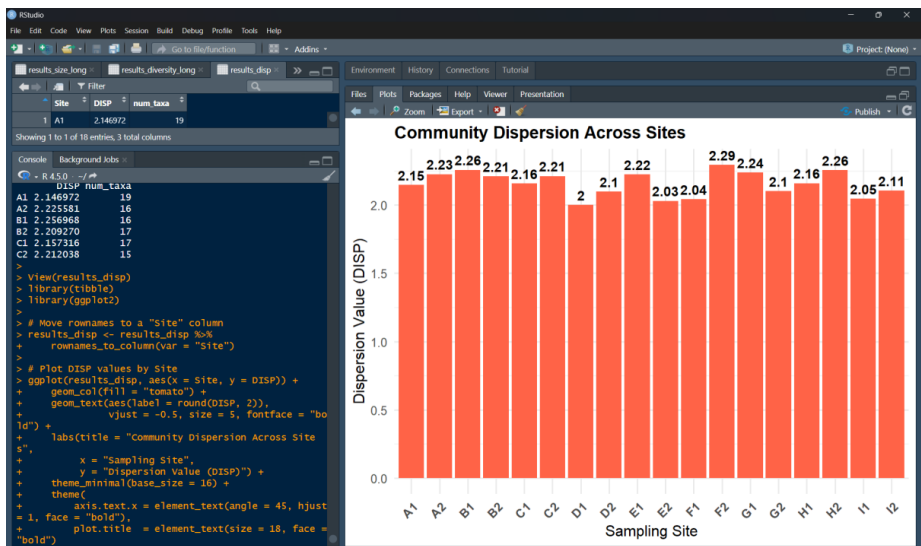


Fig. 17.26 Visualization of Dispersion (DISP) values by site generated in RStudio, used to assess variability in diatom assemblage structure.

CODE DOCUMENTATION AND EXPORT MANAGEMENT

To ensure clarity and reproducibility, it is critical to understand how this unified script manages outputs:

`diat_loadData()`: This function acts as the gatekeeper. It will scan the `input_directory` for the valid dataset. If multiple CSVs are present, it may prompt the user to specify the target file.

`diatThorAll()`: This is the powerhouse function of the package. Rather than requiring the user to type `diat_ips()`, `diat_tdi()`, and `diat_vandam()` sequentially, this command aggregates them into a single computational process.

`exportOutput = TRUE`: This is the most critical parameter for automation. By setting this argument to `TRUE`, Diathor bypasses the need to manually write export functions (like `write.csv()`). It automatically generates cleanly formatted `.csv` files for each index and trait category and deposits them directly into the defined `output_directory`.

By utilizing this streamlined script, researchers can significantly reduce code redundancy, minimize the potential for syntax errors, and ensure that their analytical workflow remains robust, rapid, and easily transferable to future datasets.

CHAPTER 19

CHEAT CODE

“Because even in science, knowing the shortcut doesn’t make you lazy – it makes you legendary.”

In computational ecology, the initial phases of learning a new software package involve executing commands step-by-step to understand the underlying mechanics of data parsing, index calculation, and visualization. However, once a researcher has mastered these foundational steps, running individual indices manually becomes inefficient, particularly when processing large-scale biomonitoring datasets across multiple projects.

The true advantage of R and the DiaThor package lies in their capacity for automation. By consolidating individual commands into a single, optimized script, researchers can transition from manual execution to automated batch processing. This chapter provides a unified, rapid-reference workflow—effectively an analytical “speedrun”—designed for power users. It allows researchers to load data, compute all major ecological indices simultaneously, and automatically export the results to a designated directory, thereby streamlining the workflow for maximum efficiency and reproducibility. While previous chapters have detailed the step-by-step execution of individual functions within DiaThor, routine biomonitoring often requires a more streamlined approach. This chapter provides a unified, automated script designed to run all applicable indices simultaneously and export the results directly into formatted CSV files. This ‘cheat code’ is particularly useful for power users processing large datasets or conducting batch analyses across multiple monitoring sites.

PREREQUISITES FOR AUTOMATED EXECUTION

Before executing the unified script, researchers must ensure that their environment and data are properly configured to prevent interruption during the automated run:

- **Data Formatting:** The input .csv file must strictly adhere to DiaThor’s formatting rules (taxa as rows, samples as columns, strictly numeric abundance values, and standard nomenclature).
- **Directory Mapping:** The working directory must be clearly defined so the script knows exactly where to pull the raw data and where to save the generated output files.
- **Library Initialization:** All necessary R packages must be loaded into the environment prior to executing the core functions.

THE UNIFIED ANALYTICAL SCRIPT:

The following code block is an optimized, all-in-one script. It executes the core diaThorAll() function, which automatically processes the dataset through every available index and functional trait database within the package.

SINGLE-SCRIPT FOR RUNNING DIATHOR INDICES AND SAVING THEM AS A .CSV FILE IN A FOLDER

Just copy paste the whole script in R Console, you have to only change the path of the folder where you wish to save your file.

```
1. Edit the line: resultsPath <- "C:/DIATHOR/CASE_2"
2. Open this script in RStudio and Run (Source) it.
3. Outputs (CSV file) will be saved in a "Diathor_results" folder.
resultsPath <- "C:/DIATHOR/CASE_2" # <---- CHANGE THIS to your case folder
output_dir <- file.path(resultsPath, "Diathor_results")
# -----
# 1. Install/load packages
# -----
needed <- c("diathor", "ggplot2", "dplyr", "tidyr", "tibble", "scales", "janitor")
for (pkg in needed) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg, dependencies = TRUE)
  }
  library(pkg, character.only = TRUE)
}
if (!dir.exists(output_dir)) dir.create(output_dir, recursive = TRUE)
# -----
```

```

# 2. Helper functions
# -----
tidy_results <- function(df, site_col = "Site") {
  if (is.null(df)) return(NULL)
  if (!site_col %in% colnames(df)) {
    df <- df %>% tibble::rownames_to_column(var = site_col)
  }
  df[[site_col]] <- as.character(df[[site_col]])
  return(df)
}

safe_run <- function(fun, df, name) {
  res <- NULL
  tryCatch({
    message("Running ", name, " ...")
    res <- fun(df)
    if (!is.null(res)) message(" -> ", name, " finished.")
  }, error = function(e) {
    message("Error in ", name, ": ", e$message)
  })
  return(res)
}

# -----
# 3. Load Diathor input
# -----
df <- NULL
tryCatch({
  df <- diat_loadData(resultsPath = resultsPath)
  message("Data loaded OK. Check 'Taxa included.csv' if warnings appear.")
}, error = function(e) {
  stop("Failed to load data: ", e$message)
})

# -----
# 4. Run all indices
# -----
modules <- list(
  IPS = function(x) diat_ips(x),
  TDI = function(x) diat_tdi(x),
  SLA = function(x) diat_sla(x),
  DDI = function(x) diat_ddi(x),
  LOBO = function(x) diat_lobo(x),
  EDI = function(x) diat_edi(x),
  ILM = function(x) diat_ilm(x),
  IDCH = function(x) diat_idch(x),
  EPID = function(x) diat_epid(x),
  IDP = function(x) diat_idp(x),
  IDAP = function(x) diat_idap(x),
  CEMFG = function(x) diat_cemfgs_rb(x),
  SPEAR = function(x) diat_spear(x),
  PBIDW = function(x) diat_pbidw(x),
  VanDam = function(x) diat_vandam(x),
  Guilds = function(x) diat_guilds(x),
  Morpho = function(x) diat_morpho(x),
  Size = function(x) diat_size(x),
  Diversity = function(x) diat_diversity(x),
  Disp = function(x) diat_disp(x)
)

```

```

)

results_list <- list()
for (nm in names(modules)) {
  res <- safe_run(modules[[nm]], df, nm)
  if (!is.null(res) && is.data.frame(res)) {
    res <- tidy_results(res)
    results_list[[nm]] <- res
    write.csv(res, file.path(output_dir, paste0("results_", nm, ".csv")), row.names =
FALSE)
  }
}

```

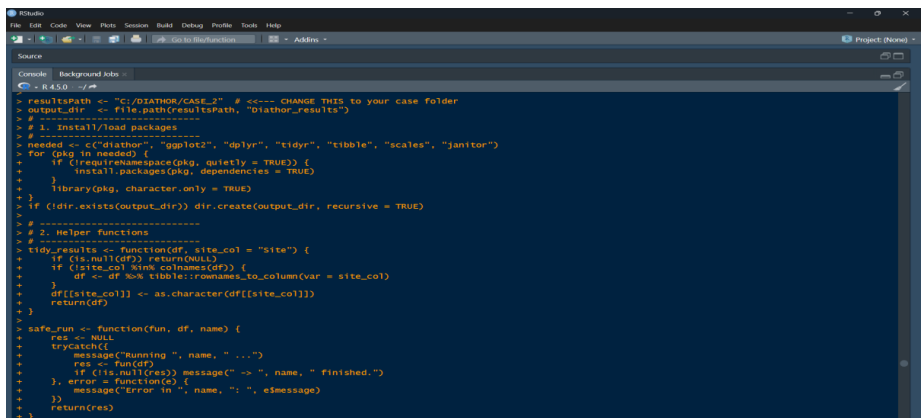
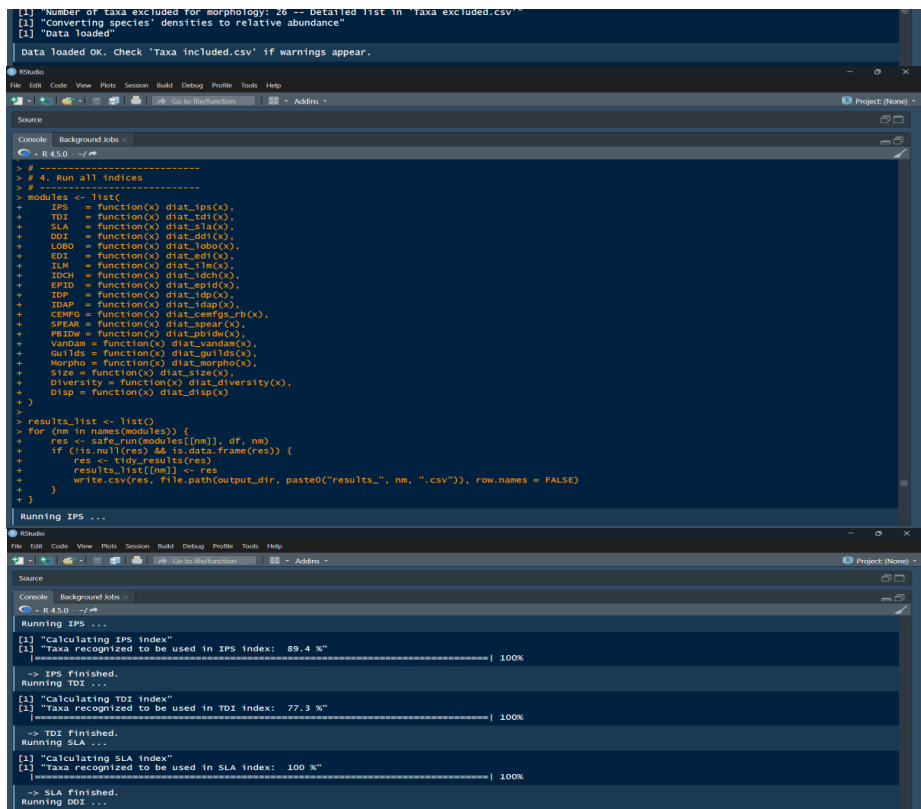


Fig. 18.1 RStudio interface showing a unified script for computing Diathor diatom indices



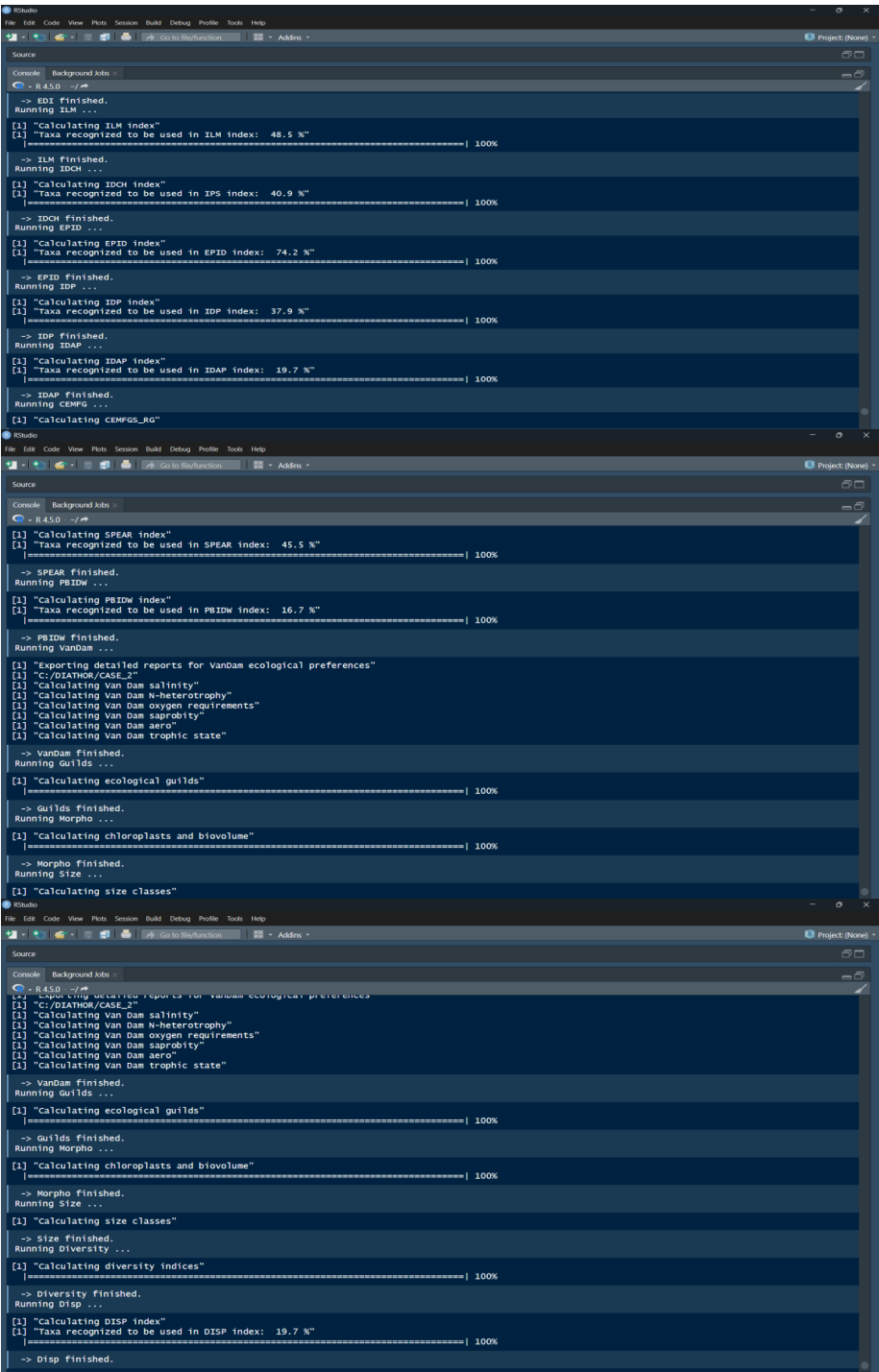


Fig. 18.3 RStudio interface showing a unified script for computing *DiaThor* diatom indices and exporting results and exporting results as CSV files.

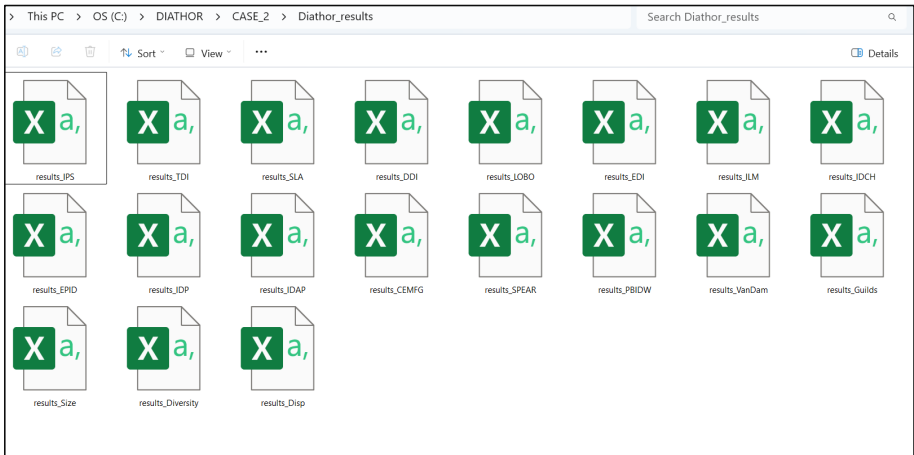


Fig. 18.4 Screenshot of exported results.

CHAPTER 20

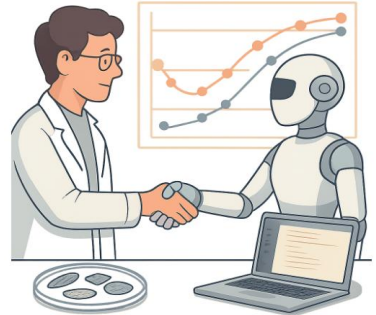
SUPER ADVANCED CSV PLOTTER WITH RECOMMENDATIONS

“When your plots start giving advice, you’ve officially crossed from analysis to alchemy.”

You know the feeling. You’ve just finished weeks of sampling in the field, trudging through wetlands, collecting bottles of water, peering at diatoms through the microscope. Finally, you sit down at your desk with a spreadsheet full of numbers: salinity values, oxygen indices, Van Dam classifications.

And then it hits you: how on earth am I going to make sense of this table?

This is where R comes in. Not the scary programming monster some people make it out to be, but a friendly tool that can help turn numbers into pictures, and pictures into stories. In this chapter, we’ll walk together through an R script designed to do exactly that; load your ecological data, ask you what kind of graph you want, and produce plots so clear that even your advisor might smile.



Basic Plotting

The core of the plotting script relies on `ggplot2` to generate foundational visualizations like scatter plots, line graphs, and bar charts based on the user's dataset.

Customization

Through interactive prompts, users can define custom titles and axis labels, avoiding the need to manually alter the code for every new dataset.

Interactivity

By integrating user prompts (e.g., `select.list()`), the script allows practitioners to choose their desired plot type dynamically. This bridges the gap between static code and interactive analysis, allowing users to rapidly pivot between visualization styles.

Saving

Once the plot is finalized, the script integrates standard export functions (e.g., `ggsave()`) to ensure the resulting graphics are saved in high-resolution formats suitable for academic publication.

Module 1: Basic Plotting

The core engine of this workflow is the basic plotting module, which reads a designated CSV file and constructs a foundational `ggplot2` object. This module establishes the data structure before any advanced features are applied.

1. Bringing Your Data Into R

The very first step is to grab your dataset. Instead of typing clunky file paths like `C:/Users/Desktop/results.csv`, we use a neat little popup:

```
file_path <- choose.files(caption = "📁 Select your CSV file")
df <- read.csv(file_path, stringsAsFactors = FALSE)
cat("\n☑ File loaded successfully!\n\n")
```

`choose.files()` → Opens a file dialog box. Think of it like the Browse button in any software.

`read.csv()` → Reads the file into R. It's like pouring your spreadsheet into a test tube labelled `df`.

The `cat("☑ ...")` line isn't necessary, but it gives us a friendly confirmation in the console.

Why does this matter? Because biology data is never static. One week you're studying salinity, the next it's dissolved oxygen. Hard-coding file paths is rigid. A file picker keeps things flexible.

Module 2: Customization

Once the basic plot is generated, researchers frequently need to tailor the visual output to specific publication standards. The customization module allows users to dynamically assign titles, adjust axis labels, and apply specific color palettes.

Module 3: Interactivity (GUI Pop-ups)

To eliminate the need for manual code editing during repetitive tasks, the interactivity module introduces graphical pop-up menus. Utilizing the `svDialogs` package, this module prompts the user to select their CSV file and preferred plot parameters dynamically.

2. Talking to the User (That's You!)

One of the biggest complaints from biologists learning R is: "I don't want to edit code every time. I just want to click and go."

That's why this script uses popup menus to ask what you want:

```
plot_type <- select.list(
  c("Scatter Plot", "Line Plot", "Bar Plot"),
  title = "☞ Choose plot type"
)
```

This shows you a popup with three options. No typing, no memorizing commands.

And what if you want to type your own labels? Here comes our helper friend `ask_text()` – a tiny function that creates a popup textbox:

```
plot_title <- ask_text("⚡ Enter your custom plot title:")
```

It's like R is politely knocking on your shoulder: "Hey, what title should I put on this graph?"

3. Numbers vs. Categories – A Subtle but Crucial Difference

Now here's a sneaky detail that trips up a lot of scientists. Look at your dataset. It probably has a column called `Site`, with values like 1, 2, 3, ...

To your brain, these are names of sampling locations. To R's brain, they are just numbers. And numbers are treated as continuous – which means R puts ticks at 5, 10, 15 instead of 1, 2, 3.

We don't want that. So, we gently remind R: "These are labels, not measurements."

```
df_long[[x_name]] <- as.factor(df_long[[x_name]])
```

This simple line forces R to treat sites as categories. Now your X-axis will proudly display 1, 2, 3, ... – exactly how you labelled your field samples.

4. The Heartbeat of the Plot – `ggplot2`

Everything we draw is built with `ggplot2`, R's graphics engine. Think of it as the microscope of data visualization – powerful, flexible, but only if you learn how to adjust the lenses.

The basic call looks like this:

```
p <- ggplot(df_long, aes(x = Site, y = value, color = variable))
```

This line says:

Take the data `df_long` (our dataset in “long” format).

Put Site on the X-axis.

Put value (e.g., salinity or oxygen) on the Y-axis.

Use different colours for different variables.

After this, the magic begins. We add layers, just like adding stains to a microscope slide.

Scatter Plot

```
p <- p + geom_point(size = 3) +  
  geom_text(aes(label = round(value, 2)), vjust = -0.7, size = 3)
```

Little dots appear for each observation. And see that `geom_text()`? That's what puts the numbers above the dots, so you don't have to squint back and forth between graph and table.

Line Plot

```
p <- p + geom_line(size = 1) + geom_point(size = 2)
```

Perfect when you want to see trends across sites. For example, how oxygen changes from upstream to downstream.

Bar Plot

```
p <- p + geom_bar(stat = "identity", position = "dodge") +  
  geom_text(aes(label = round(value, 2)), vjust = -0.5, size = 3)
```

Bars are chunky and clear, excellent for comparing totals across sites. And again, the numbers sit neatly above the bars.

5. Dressing Up the Graph

Just like you wouldn't present at a conference in muddy field clothes (well, usually), your plots deserve a clean look. That's why we let you pick a theme:

```
if (theme_choice == "Classic") p <- p + theme_classic()  
if (theme_choice == "Minimal") p <- p + theme_minimal()  
if (theme_choice == "Dark")    p <- p + theme_dark()
```

Classic: White background, sharp axes.

Minimal: Stripped down, light, and modern.

Dark: Dark background, glowing points. Perfect for posters.

Changing themes is like changing the mood of your story.

Purpose: This makes the R environment significantly more accessible, allowing users to process data rapidly using familiar point-and-click interfaces rather than rewriting variable names.

Module 4: Saving and Exporting

The final module automates the preservation of the generated visuals. It takes the finalized ggplot2 object and exports it as a high-resolution image suitable for academic publication.

6. Saving for Later

The final step is preserving your work:

```
ggsave("plot.png", p, width = 8, height = 6, dpi = 300)
```

width and height set the size in inches.

dpi = 300 gives publication quality – sharp enough for journals.

Think of this as bottling your graph like a water sample – sealed, labelled, and ready to ship.

7. Why Biologists Should Care

Okay, so we've built a script. But why does it matter? Here's the bigger picture:

- **Clarity:** Numbers in tables are abstract. Plots let your audience see patterns – rising salinity, falling oxygen.
- **Flexibility:** You don't have to rewrite code each time. Just pick your plot and labels in popups.
- **Reproducibility:** Tomorrow, when you get new samples, you can re-run the exact same script. No reinventing the wheel.
- **Professional polish:** Auto-labels and themes save you from clunky Excel charts.

Most importantly, it frees your brain from formatting battles so you can focus on ecological interpretation.

8. From Numbers to Narratives

Science isn't just about data collection. It's about storytelling. This script helps you tell the story of your river, your diatoms, your oxygen levels – clearly and visually.

Instead of a static spreadsheet, you now have living, breathing plots. And all it took was a bit of code, written in a way that speaks your language.

THE UNIFIED EXECUTION PIPELINE

By calling these modules sequentially, the entire interactive workflow operates seamlessly:

#CODE:

```
# =====
# Interactive Plot Builder with GUI popups
# (All inputs via dialogs, no console typing)
# =====

library(ggplot2)
library(plotly)
library(reshape2)
```

```

library(tcltk)

# --- Helper function for popup text input ---
ask_text <- function(prompt = "Enter text:", default = "") {
  tt <- tktoplevel()
  tkwm.title(tt, "Custom Input")

  text_var <- tclVar(default)
  lbl <- tklabel(tt, text = prompt)
  ent <- tkentry(tt, textvariable = text_var, width = 40)
  ok <- tkbutton(tt, text = "OK", command = function() tkdestroy(tt))

  tkpack(lbl, ent, ok, padx = 10, pady = 5)
  tkwait.window(tt)

  return(tclvalue(text_var))
}

# 1. Load CSV
file_path <- choose.files(caption = "📁 Select your CSV file")
df <- read.csv(file_path, stringsAsFactors = FALSE)

cat("\n✅ File loaded successfully!\n\n")
print(head(df))
cat("\n🔍 Column names:", names(df), "\n")

# 2. Select X
cat("\n💡 Recommendation: If your X is categorical (like Site, Species),
  Bar/Box plots work best. If numeric (like Time, Concentration),
  Line/Scatter plots are more suitable.\n")
x_name <- select.list(names(df), title = "👉 Select column for X-axis", graphics = TRUE)

# 3. Select Y
cat("\n💡 Recommendation:
  - Single Y → simpler plots (Bar, Scatter).
  - Multiple Y → use Line, Scatter, or grouped Bar plots for comparison.\n")
y_names <- select.list(names(df), title = "👉 Select column(s) for Y-axis (tick multiple
if needed)",
                      graphics = TRUE, multiple = TRUE)
if (length(y_names) == 0) stop("❌ You must select at least one Y column!")

# 4. Smart Plot Suggestion
valid_plot_types <- c()
if (is.character(df[[x_name]]) | is.factor(df[[x_name]])) {
  valid_plot_types <- c("Bar Plot", "Box Plot")
} else {
  valid_plot_types <- c("Scatter Plot", "Line Plot", "Bar Plot", "Area Plot")
}
cat("\n💡 Recommendation: Based on your X selection ('", x_name, "'),
  the best suited plot types are:", paste(valid_plot_types, collapse = ", "), "\n")
plot_type <- select.list(valid_plot_types, title = "👉 Choose suitable plot type",
graphics = TRUE)

# 5. Faceting
cat("\n💡 Recommendation: Faceting is useful if you want to split data by groups
(e.g., Year, Site Type, Season).\n")
facet_choice <- select.list(c("No", "Yes"), title = "👉 Do you want to facet the plot?",
graphics = TRUE)

```



```

facet_col <- NULL
if (facet_choice == "Yes") {
  facet_col <- select.list(names(df), title = "👉 Select faceting column", graphics =
TRUE)
}

# 6. Customizations
cat("\n👉 Tip: Choose auto color if plotting multiple Y variables to auto-assign distinct
colors.\n")
plot_color <- select.list(c("blue", "red", "forestgreen", "orange", "purple", "black",
"auto"),
                          title = "👉 Choose plot color", graphics = TRUE)

cat("\n👉 Tip:
- Minimal = clean, publication ready
- Classic = traditional axes
- Light = soft background
- Dark = dramatic effect\n")
theme_choice <- select.list(c("Minimal", "Classic", "Light", "Dark"),
                           title = "👉 Choose ggplot theme", graphics = TRUE)

plot_title <- select.list(c("My Plot", "Index by Site", "Custom Title"),
                         title = "👉 Choose plot title", graphics = TRUE)
if (plot_title == "Custom Title") {
  plot_title <- ask_text("👉 Enter your custom plot title:")
}

x_label <- select.list(c(x_name, "Custom X Label"), title = "👉 Choose X-axis label",
graphics = TRUE)
if (x_label == "Custom X Label") x_label <- ask_text("👉 Enter custom X-axis label:")

y_label <- paste(y_names, collapse = ", ")
y_label <- select.list(c(y_label, "Custom Y Label"), title = "👉 Choose Y-axis label",
graphics = TRUE)
if (y_label == "Custom Y Label") y_label <- ask_text("👉 Enter custom Y-axis label:")

title_size <- as.numeric(select.list(c("14", "16", "18", "20", "22"),
                                     title = "👉 Title font size", graphics = TRUE))
axis_size <- as.numeric(select.list(c("10", "12", "14", "16", "18"),
                                     title = "👉 Axis label font size", graphics = TRUE))

legend_pos <- select.list(c("right", "left", "top", "bottom", "none"),
                        title = "👉 Choose legend position", graphics = TRUE)

show_labels <- select.list(c("Yes", "No"), title = "👉 Show values above bars/points?",
graphics = TRUE)

# 7. Build Plot
df_long <- melt(df, id.vars = x_name, measure.vars = y_names)

# ☒ Force X-axis to be categorical (for proper Site labels)
df_long[[x_name]] <- as.factor(df_long[[x_name]])

p <- ggplot(df_long, aes_string(x = x_name, y = "value", color = "variable", fill =
"variable"))

if (plot_type == "Bar Plot") {
  p <- p + geom_col(position = "dodge", alpha = 0.8)
}

```

```

    if (show_labels == "Yes") p <- p + geom_text(aes(label = round(value, 2)),
                                                position = position_dodge(width = 0.9),
vjust = -0.3)
  } else if (plot_type == "Line Plot") {
    p <- p + geom_line(size = 1.2) + geom_point(size = 3)
    if (show_labels == "Yes") p <- p + geom_text(aes(label = round(value, 2)), vjust = -
0.7, size = 3)
  } else if (plot_type == "Scatter Plot") {
    p <- p + geom_point(size = 3)
    if (show_labels == "Yes") p <- p + geom_text(aes(label = round(value, 2)), vjust = -
0.7, size = 3)
  } else if (plot_type == "Box Plot") {
    p <- p + geom_boxplot()
  } else if (plot_type == "Area Plot") {
    p <- p + geom_area(position = "stack", alpha = 0.6)
  }

  if (facet_choice == "Yes" & !is.null(facet_col)) p <- p + facet_wrap(as.formula(paste("~",
facet_col)))

  if (theme_choice == "Minimal") p <- p + theme_minimal()
  if (theme_choice == "Classic") p <- p + theme_classic()
  if (theme_choice == "Light") p <- p + theme_light()
  if (theme_choice == "Dark") p <- p + theme_dark()

  p <- p + labs(title = plot_title, x = x_label, y = y_label) +
    theme(plot.title = element_text(size = title_size, face = "bold", color = "darkblue",
hjust = 0.5),
          axis.title.x = element_text(size = axis_size, face = "bold"),
          axis.title.y = element_text(size = axis_size, face = "bold"),
          axis.text = element_text(size = 12),
          legend.position = legend_pos)

  print(p)

# 8. Interactive
cat("\n🔗 Recommendation: Interactive plots are useful for exploring data (zoom, hover).
  But for publication-quality figures, static ggplot is preferred.\n")
interactive_choice <- select.list(c("No", "Yes"), title = "🔗 Do you want an interactive
plot (Plotly)?", graphics = TRUE)
if (interactive_choice == "Yes") {
  p <- ggplotly(p)
  print(p)
}

# 9. Save Options
cat("\n🔗 Recommendation:
  - PNG/JPEG = presentations
  - PDF/SVG = vector graphics for publications
  - TIFF = high DPI journals\n")
save_plot <- select.list(c("Yes", "No"), title = "🔗 Do you want to save the plot?",
graphics = TRUE)
if (save_plot == "Yes") {
  folder <- tk_choose.dir(caption = "📁 Select folder to save plot")
  format_choice <- select.list(c("PNG", "JPEG", "TIFF", "PDF", "SVG"),
                              title = "🔗 Choose file format", graphics = TRUE)

  width <- as.numeric(ask_text("📏 Enter width in inches:", "10"))

```

```

height <- as.numeric(ask_text("Enter height in inches:", "6"))
dpi <- as.numeric(ask_text("Enter resolution DPI:", "300"))

file_name <- ask_text("Enter filename (without extension):", "my_plot")
file_path <- file.path(folder, paste0(file_name, ".", tolower(format_choice)))

ggsave(file_path, p, width = width, height = height, dpi = dpi, device =
tolower(format_choice))
cat("\nPlot saved as", file_path, "\n")
}

cat("\nDone! Your interactive plot is ready with recommendations at every step!\n")
width_input <- readline("Enter width in inches (default 10): ")
if (width_input == "") width <- 10 else width <- as.numeric(width_input)

height_input <- readline("Enter height in inches (default 6): ")
if (height_input == "") height <- 6 else height <- as.numeric(height_input)

dpi_input <- readline("Enter resolution DPI (default 300): ")
if (dpi_input == "") dpi <- 300 else dpi <- as.numeric(dpi_input)

file_name <- readline("Enter filename (without extension): ")
file_path <- file.path(folder, paste0(file_name, ".", tolower(format_choice)))

ggsave(file_path, p, width = width, height = height, dpi = dpi, device =
tolower(format_choice))
cat("\nPlot saved as", file_path, "\n")
}

cat("\nDone! Your interactive plot is ready with recommendations at every step!\n")

```

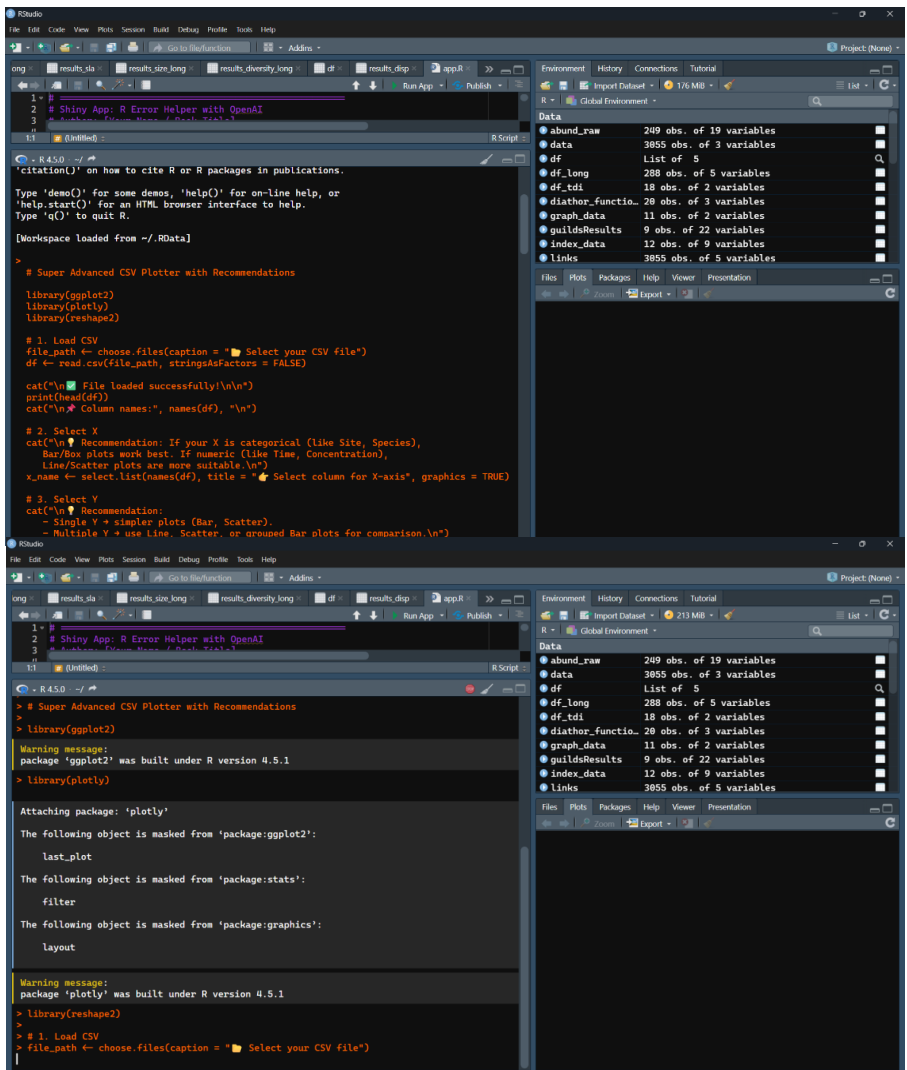


Fig. 19.1 RStudio interface of a super-advanced CSV plotter with automated visualization and recommendation features.

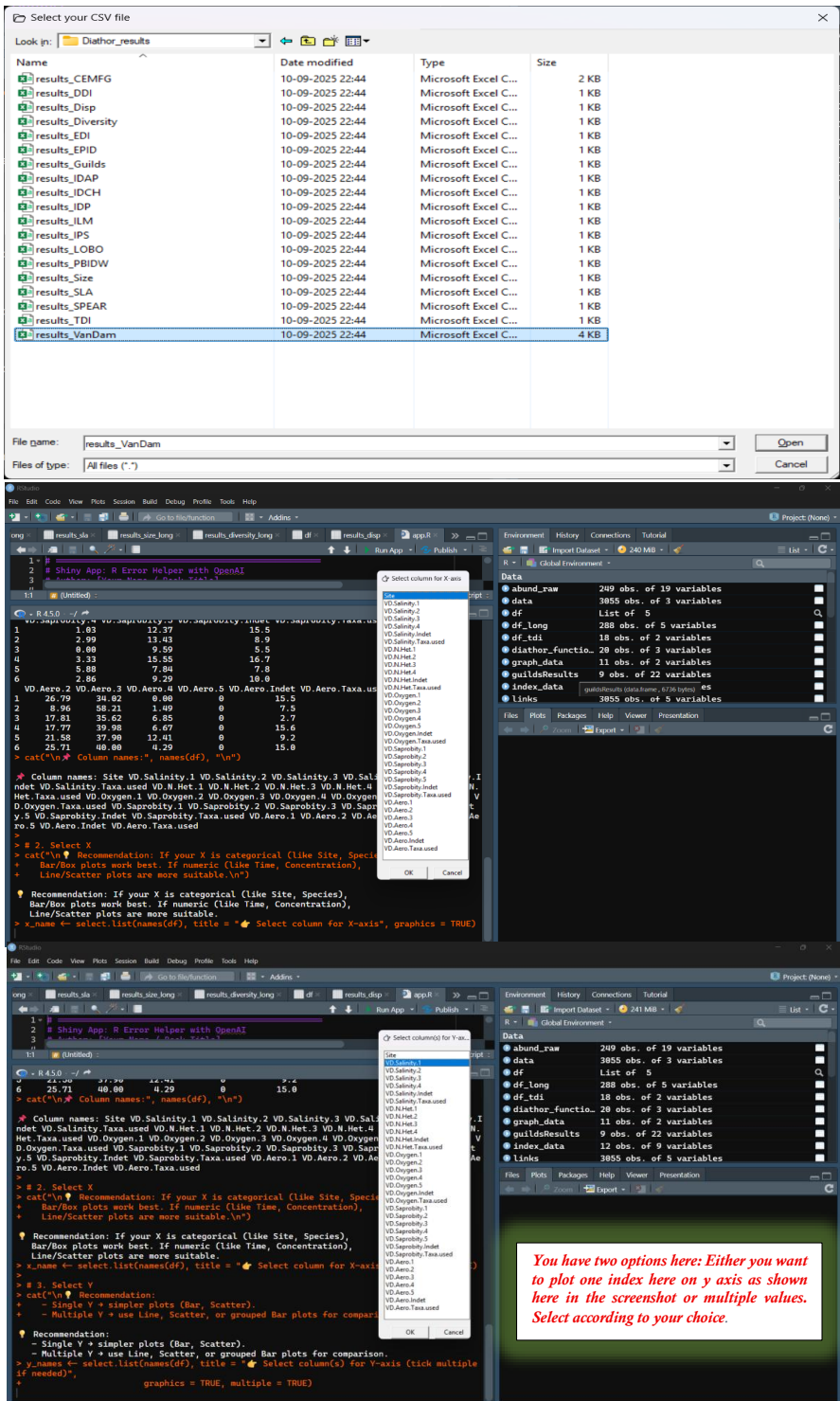


Fig. 19.2 RStudio interface of a super-advanced CSV plotter with automated visualization and recommendation features.

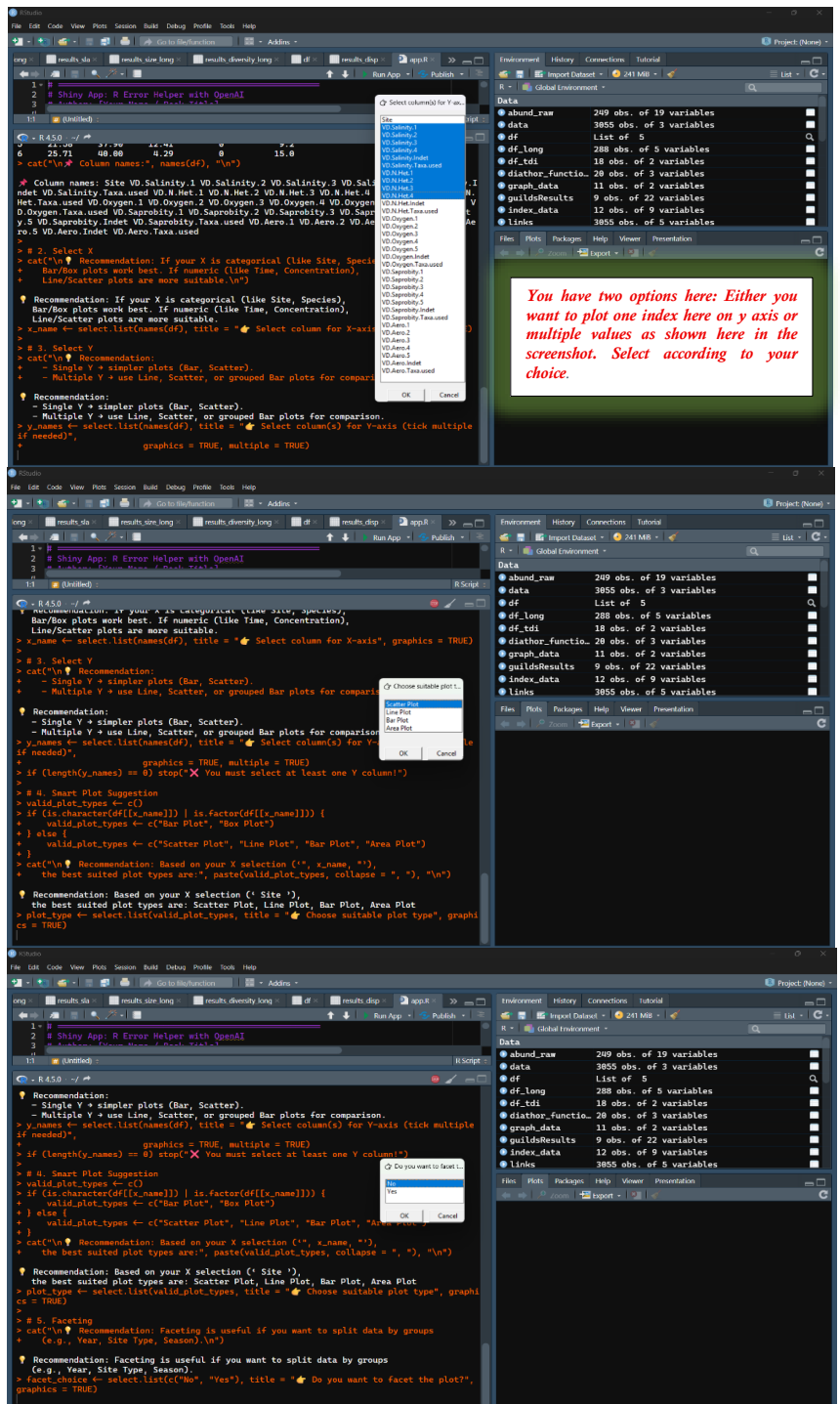


Fig. 19.3 RStudio interface of a super-advanced CSV plotter with automated visualization and recommendation features.

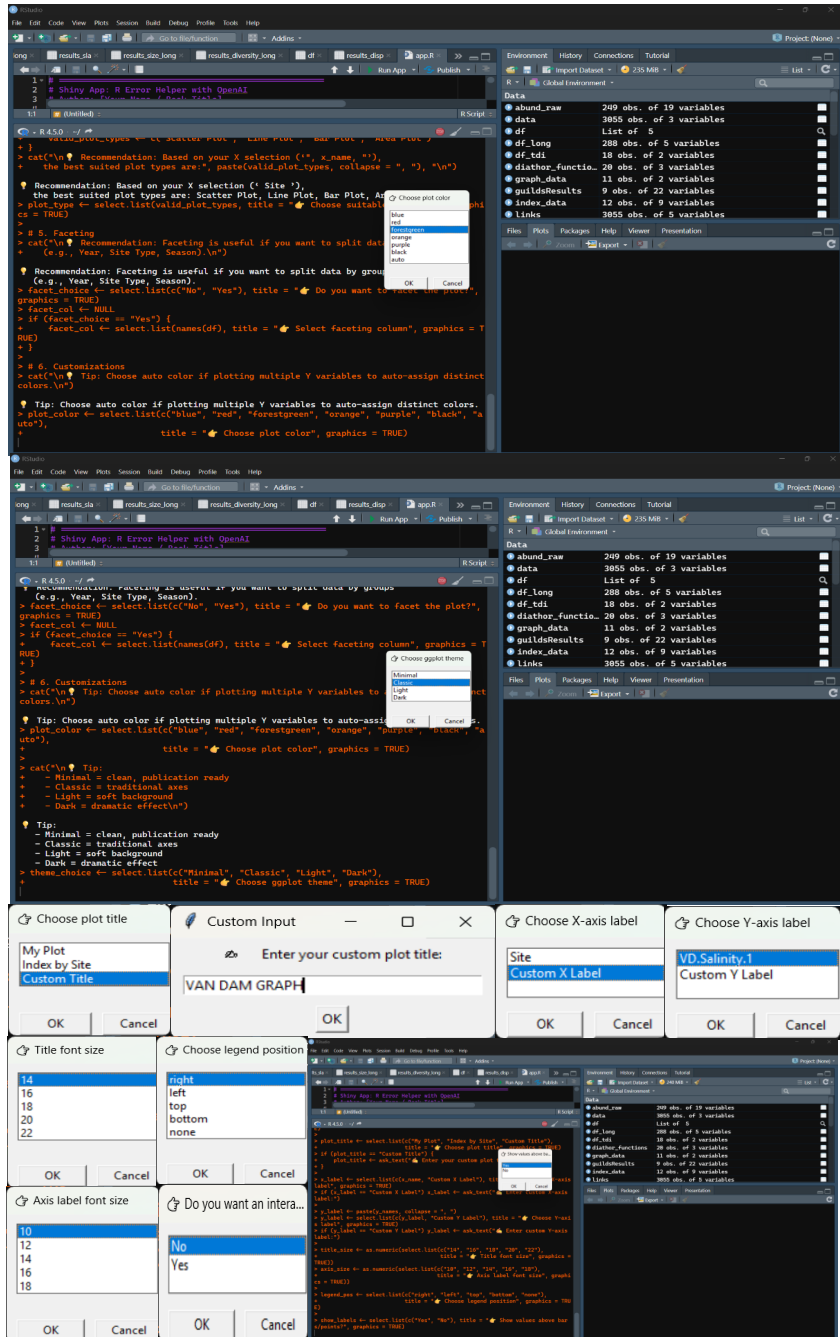


Fig. 19.4 RStudio interface screenshot of prompts of a super-advanced CSV plotter with automated visualization and recommendation features.

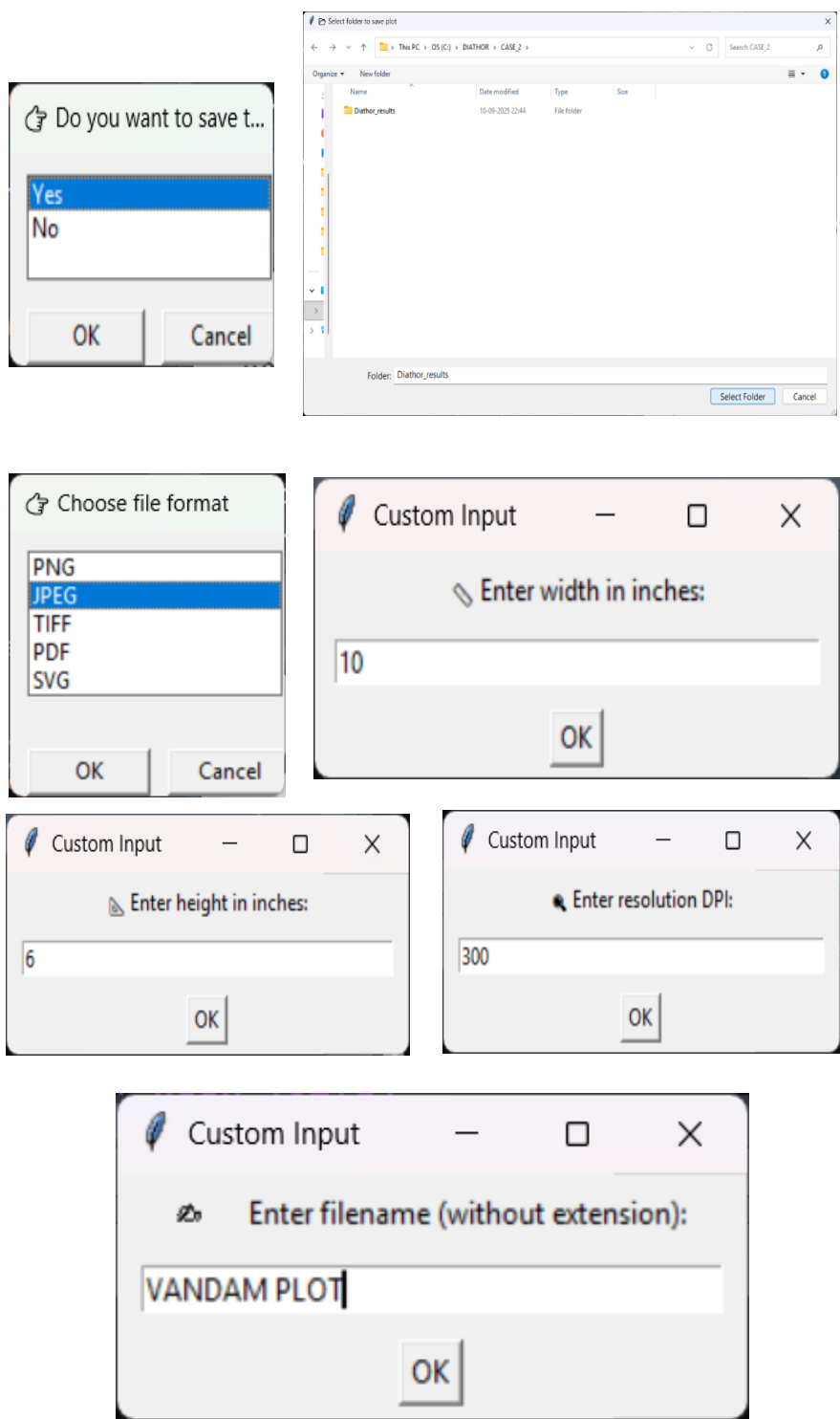


Fig. 19.5 RStudio interface screenshot of prompts of a super-advanced CSV plotter with automated visualization and recommendation features.

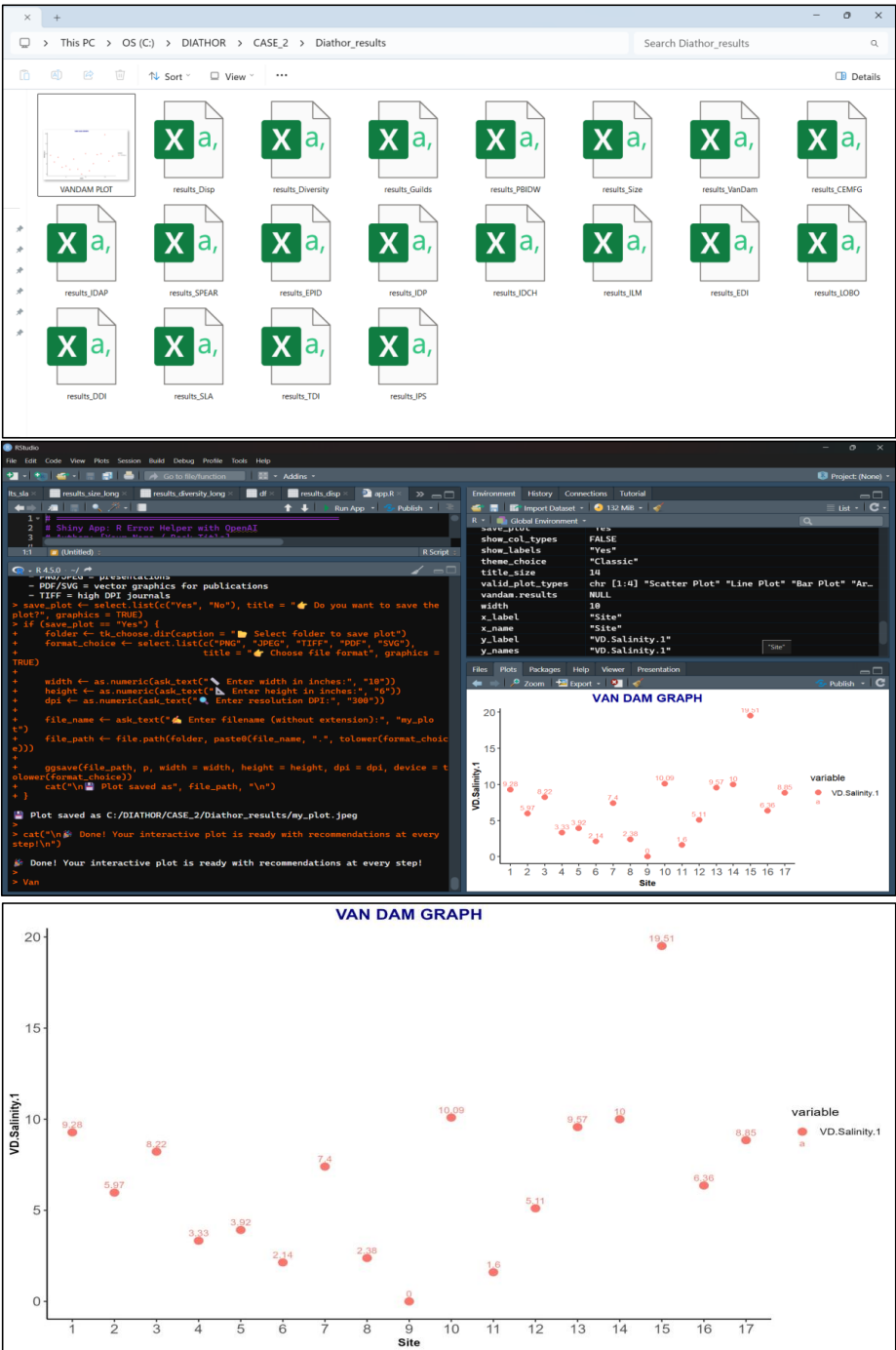


Fig. 19.6 RStudio interface screenshot of prompts of a super-advanced CSV plotter with automated visualization and recommendation features.

CHAPTER 21

HOW USEFUL IS DIATHOR PACKAGE?

"A tool's worth isn't in its code – it's in how clearly it helps you see what nature's been saying all along."

Everyone in ecological data analysis eventually hits that phase where they've got raw diatom data, thousands of species names, endless columns of counts and a single question echoing in their brain: "Now what?"

Enter DiaThor, the R package that basically says, "Relax, I've got this."

It takes your messy species abundance data, waves its algorithmic wand, and — bam! — out come dozens of ecological indices. IPS, TDI, GDI, IBD, SLA... all those cryptic acronyms suddenly start meaning something.

But the real question isn't what it does — it's how useful it actually is.

Let's unpack that like we're over coffee and your laptop's open.

So, what makes DiaThor actually useful?

a) It saves weeks of manual work.

Back in the day, calculating indices meant juggling Excel sheets, copying formulas, referencing diatom databases, and hoping you didn't misspell *Navicula radiosa*. DiaThor automates all that. It recognizes taxa (when they're spelled right), fetches the corresponding ecological values, and computes indices in seconds. That's not just efficiency — that's sanity-saving automation.

b) It standardizes your analysis.

Reproducibility is the holy grail of science. DiaThor uses recognized European standards and established formulae for each index, so your results can actually be compared across studies and time. In short — your data doesn't just stay in your lab; it can talk to the world's data.

c) It's transparent.

Unlike black-box software, every step in DiaThor is traceable. You can open the source code, see how indices are computed, and even tweak formulas if you're brave enough. It's open-source — science's way of saying, "Trust, but verify."

d) It plays nice with R.

This isn't a closed ecosystem. DiaThor outputs clean CSVs, meaning you can instantly visualize, filter, or integrate them with other tools, ggplot2, Shiny, or even custom scripts. It's like having a base station for ecological analysis that plugs into any data pipeline. But here's the reality check. No tool is flawless and DiaThor has its growing pains.

Species name mismatches can trip it up. Misspell a taxon, and the software stares blankly at you like, "Never heard of her."

The database coverage is good but not universal — especially for lesser-known tropical taxa.

And yes, there's a learning curve if you're not familiar with R.

But honestly, those "flaws" are just reminders that the software is built for scientists, not consumers. It expects you to think, not just click.

So how useful is it, really?

If you're an ecologist, a researcher, or even a student diving into bioassessment, it's a game-changer.

DiaThor isn't just a tool; it's an accelerator for understanding. It transforms data into ecological stories — about pollution, resilience, and recovery.

It's useful not because it replaces thinking, but because it frees you to think about what truly matters: what the diatoms are saying about your waters, your ecosystems, and, indirectly, your future.

WHY THE DIATHOR PACKAGE IS

Useful

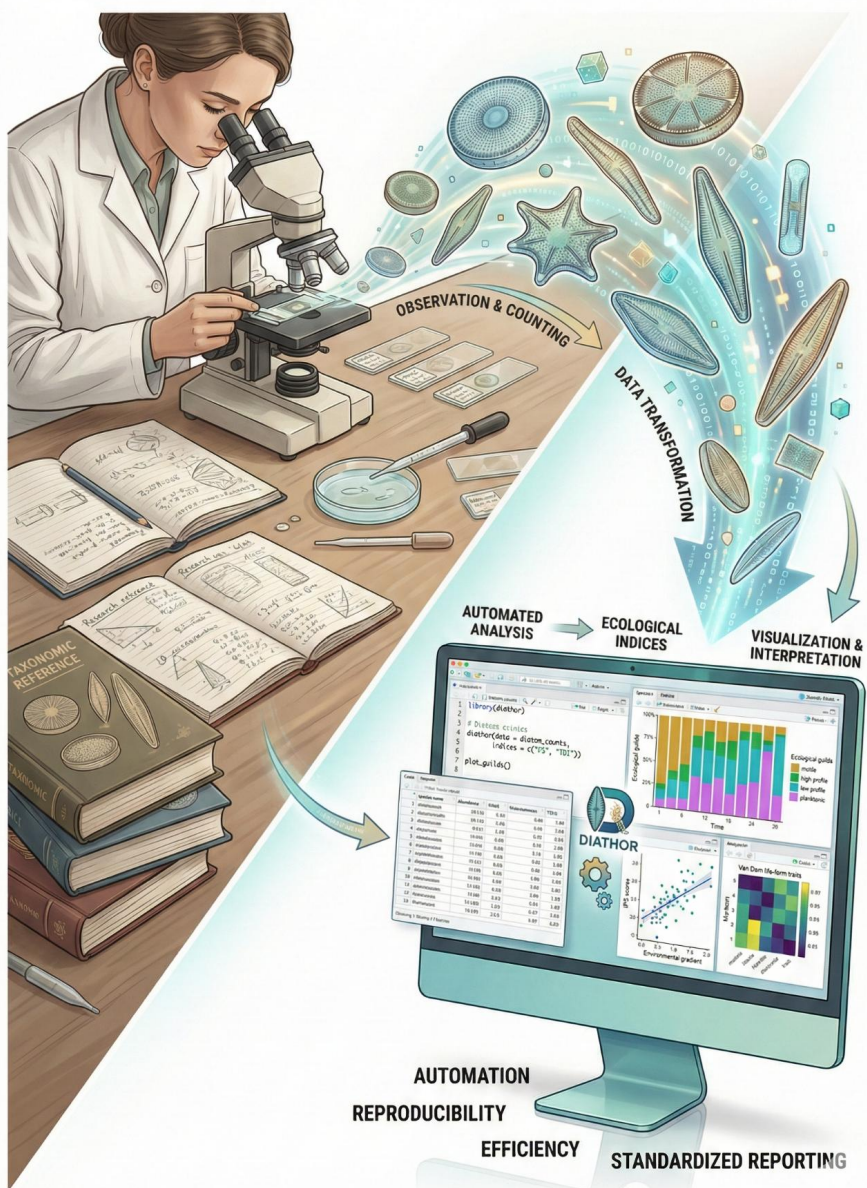
- Calculates dozens of diatom indices (IPS, TDI, GDI, IBD, etc.) automatically.
- Converts raw abundance data into meaningful ecological scores in seconds.
- Eliminates tedious manual index calculations and Excel chaos.
- Ensures consistency and reproducibility across studies.
- Uses standardized international formulas and references (e.g., OMNIDIA-style indices).
- Integrates multiple ecological parameters – saprobity, trophic state, salinity, and more.
- Generates clean, ready-to-use CSV outputs for further analysis.
- Offers built-in plotting options for quick visualization of results.
- Provides clear taxon matching and warns about unrecognized species.
- Encourages open, transparent, and auditable science.
- Works smoothly with Shiny apps for interactive dashboards.
- Saves time, energy, and the occasional existential breakdown before deadlines.
- Supports integration with other R packages (ggplot2, dplyr, tidyr).
- Enables multi-site comparisons effortlessly.
- Bridges biology with data science – turning field samples into statistical insight.

WHY LEARNING RSTUDIO (AND BASIC CODING)

Matters

- Lets you handle massive ecological datasets without crashing your laptop.
- Frees you from Excel dependency and copy-paste chaos.
- Automates repetitive tasks – one script, infinite reuse.
- Creates publication-quality plots and figures with precision control.
- Offers complete transparency: you see how your data is processed.
- Encourages reproducible research – anyone can rerun your analysis.
- Gives access to hundreds of ecological and statistical packages.
- Builds confidence to question software outputs rather than blindly trust them.
- Turns data cleaning from a nightmare into a neat workflow.
- Opens doors to data storytelling – code, visualize, and communicate like a pro. Strengthens analytical thinking – coding trains your brain to see patterns and question assumptions.
- Gives you control over your workflow – no more “black box” software where you don’t know what’s happening behind the scenes.
- Future-proofs your career – data literacy is fast becoming as essential as knowing how to use a microscope.

DIATHOR R PACKAGE: AUTOMATING ECOLOGICAL & DIATOM ANALYSIS



In short:

DiaThor doesn't just calculate indices.

It translates life's complexity into numbers you can work with.

And that — in the world of environmental science — is pure gold. The DiaThor package includes a comprehensive suite of established diatom-based indices from around the world. For example, each index has its own formula and

scale, but all serve to translate diatom community data into an environmental quality indicator. The broad coverage of indices in DiaThor means users can calculate multiple indices on the same dataset to get a more complete picture of water quality.

As an open-source R package, DiaThor reduces barriers to diatom analysis making advanced ecological assessments accessible to researchers and water managers particularly in resource- limited regions. Its modular design encourages continuous improvement of new indices, regional calibrations, and trait expansions can be incorporated, keeping the platform current and globally relevant.

Looking ahead, DiaThor has the potential to evolve when to a more adaptive and user-friendly tool. Future development might include taxonomic harmonization using live links to Diat.barcode or Algalbase, machine learning to predict missing traits, special mapping integration via packages like leaflet or sf, a graphical user interface (GUI) or Shiny app for broader accessibility, expansion of the internal trait database to better represent tax are from across the globe, and a community driven depository of case studies and feedback to enhance user learning and tool refinement.

In summary, DiaThor exemplifies the synergy between ecological expertise and open-source computation. It transforms raw biological data into actionable insights enabling accurate reproducible assessments of riverine health. As global water quality challenges intensify, tools like DiaThor will play a pivotal role in safeguarding aquatic ecosystems through science-based monitoring and management. The advancement of ecological modelling depends not only on biological insight but on the ingenuity of data scientists and engineers who translate complex ecosystems into code; build tools like DiaThor, and empower us to monitor nature with clarity, consistency, and scale, that is transforming diatom data into code & algorithms to build dynamic models, scaling conservation, with fusion of ecology, data engineering and machine-learning, because safeguarding water isn't just research, it's a global responsibility.

CHAPTER 22

LIMITATIONS OF USING DIATHOR AND ITS PLOTTING FUNCTIONS

"Even the sharpest tool can falter when the data dulls its edge – limitations aren't flaws; they're reminders to stay curious."

Every scientific tool, no matter how sophisticated, carries its own set of constraints. The DiaThor package, despite being a powerful ally for diatom-based ecological assessment, is no exception. Its efficiency, accuracy, and usability are tightly bound to the quality of its internal databases, user skill, and compatibility with other R packages. The following discussion unpacks these limitations, not to discredit the tool, but to help users navigate its practical boundaries with clarity and confidence.

1) Taxonomic Constraints

A recurring challenge arises when species names in the user's dataset don't align with the reference database. Since DiaThor relies on strict taxonomic matching, any discrepancy, be it a synonym, misspelling, or outdated classification — results in missing index values. In ecological research, even a few unrecognized taxa can distort overall assessments. Thus, careful pre-validation of species names is indispensable before running analyses.

2) Index Coverage Bias

DiaThor incorporates a substantial but incomplete set of ecological indices. While it effectively computes several classical indices such as IPS, TDI, and IDSE, others — particularly regional or emerging metrics — are absent. This selective inclusion introduces bias and may limit the interpretative depth of multi-index comparisons. Users must therefore be aware that DiaThor offers a representative, not exhaustive, suite of metrics.

3) Need for Continuous Taxonomic Updates

Taxonomy evolves continuously, with diatom classifications being revised almost annually. The database embedded within DiaThor must therefore be periodically updated to reflect these changes. If updates lag, the system risks obsolescence, leading to misidentification or omission of newly recognized species. This limitation underscores the need for a community-driven, dynamically maintained reference structure.

4) Limited Plot Flexibility

The visualization tools included in DiaThor serve a basic purpose: they summarize index results quickly and clearly. However, these plots are static, minimally customizable, and often aesthetically plain. Researchers seeking publication-quality graphics must export the data and reconstruct plots manually using ggplot2 or similar libraries. This reduces convenience, though it simultaneously encourages greater user creativity.

5) Steep Learning Curve

DiaThor, like most R-based tools, assumes a certain level of programming literacy. Beginners unfamiliar with R syntax, data structuring, or file management may find the initial setup intimidating. There is no graphical interface to ease this process, and mistakes in data formatting can halt analyses entirely. The learning curve, though steep, ultimately sharpens one's understanding of ecological data processing — but it demands patience and persistence.

6) Dependence on the Reference Database

Every result DiaThor produces is only as strong as its reference database. If that database lacks certain taxa, traits, or ecological tolerances, the analysis will inevitably inherit those gaps. This dependency on backend completeness is both a strength (centralized accuracy) and a vulnerability (restricted scope). The quality of the database effectively defines the boundaries of interpretation.

7) Static Nature of Plots

The package's plots are static by design; they provide snapshots rather than interactive environments. Users cannot zoom, hover, or filter data directly within the plots. For dynamic exploration or educational demonstrations, integration with Shiny, Plotly, or other interactive frameworks is required. While this adds flexibility, it also demands additional coding expertise.

8) Handling of Rare Species

Rare taxa often act as ecological indicators, yet DiaThor may misrepresent or exclude them due to limited tolerance data. This oversight can diminish the ecological nuance of the analysis. Researchers focusing on sensitive or pristine environments must therefore interpret outputs with caution, ensuring that rare species are properly reviewed outside the automated pipeline.

9) Data Preprocessing Requirements

DiaThor's input system is highly sensitive to formatting. The CSV file must conform to specific column names, species arrangements, and numeric formats. Even minor inconsistencies, such as an extra space, a missing value, or a lowercase species name, can lead to errors or null results. As a result, meticulous data cleaning and standardization are non-negotiable prerequisites.

10) Limited Error Feedback

When something goes wrong, DiaThor's error messages are often brief, vague, or entirely absent. This can leave users puzzled, particularly when dealing with complex datasets. Without explicit feedback, troubleshooting becomes a game of trial and error. The lack of descriptive diagnostics reduces efficiency and can frustrate even seasoned R users.

11) Absence of Real-time Data Validation

DiaThor processes data in batch mode, which means that errors are detected only after computation begins. There is no built-in mechanism to validate input data in real time. Consequently, users may only learn of data inconsistencies after a long computation has failed, wasting both time and effort.

12) Memory and Performance Constraints

R is known to struggle with large datasets, and DiaThor is no exception. When dealing with extensive temporal or spatial data matrices, the program may slow down or even crash due to memory limits. This restricts its scalability for regional or continental-scale studies unless data are processed in subsets.

13) Lack of Multi-language Support

At present, all functions, messages, and documentation are provided exclusively in English. This limits accessibility for non-English-speaking researchers and students, especially in regions where diatom monitoring is gaining momentum. A multilingual interface would significantly enhance global usability.

14) Limited Integration of Functional Traits

Functional trait analyses — crucial for understanding diatom ecology — remain underdeveloped within DiaThor. While the package provides basic ecological guild classifications, more sophisticated trait-based indices must be computed manually or through external packages. This limits the package's role in advanced community ecology studies.

15) Plot Label Overlaps

Graphical outputs, particularly network and bubble plots, often suffer from overlapping labels. This visual clutter reduces readability and interpretability. Adjusting label positions or font sizes usually requires manual post-processing, adding extra work for visualization refinement.

16) Absence of Built-in Interactive Dashboards

Although DiaThor generates numerous indices, it offers no built-in mechanism to explore them interactively. Users desiring a dashboard-style experience — where indices can be filtered, compared, or visualized dynamically — must construct custom Shiny applications. This introduces a technical barrier but opens a path toward high-level interactivity.

17) Limited Scope for Custom Statistics

The package does not allow users to define new indices or modify formulas within the interface. Extending its statistical capability requires direct coding and understanding of its internal structure. In essence, it caters to users who are comfortable modifying source code — a limitation for those expecting a “plug-and-play” approach.

18) Dependency on Other R Packages

DiaThor depends on multiple external libraries such as ggplot2, igraph, and dplyr. Version mismatches or updates in these dependencies can sometimes break functions or alter output behavior. This interconnectedness is typical of R ecosystems but does require careful environment management.

19) Lack of Native Time-Series Visualization

Ecological monitoring often demands temporal trend analysis, yet DiaThor is not optimized for time-series visualization. While users can manually restructure data for such plots, no built-in function exists to display how indices evolve over months or years. This leaves a methodological gap for long-term ecological assessments.

20) Incomplete Documentation for Advanced Features

Although basic usage is well-explained, advanced functionalities — such as parameter tuning or data merging — lack detailed documentation. This opacity can discourage exploration and slow down research workflows. Comprehensive user guides or community-driven wikis could greatly enhance the learning experience.

21) Limited Support for Multi-site Comparisons

When working with numerous sites, researchers often need comparative summaries. DiaThor, however, handles one dataset at a time without native cross-site aggregation functions. Users must manually reshape and merge outputs, which can be tedious for large-scale biomonitoring programs.

22) No Automatic Data Cleaning

Data integrity is assumed, not ensured. DiaThor will not correct typographical errors, remove duplicates, or fill missing entries automatically. Such preprocessing remains the user's responsibility. While this maintains analytical rigor, it also burdens the workflow with repetitive quality checks.

23) Difficulty Integrating with Other Software

Exporting results from DiaThor into GIS software or spreadsheet applications is possible but not seamless. File conversion, formatting adjustments, and manual verification steps are required. Direct integration pipelines would streamline this process and promote cross-platform ecological analysis.

24) No Built-in Confidence or Sensitivity Analysis

DiaThor reports numerical index values without expressing their uncertainty or sensitivity. Researchers must calculate confidence intervals or perform sensitivity analyses independently. The absence of uncertainty quantification limits interpretative robustness, particularly for management or policy-oriented studies.

25) Limited Graphical Customization

Finally, while DiaThor's visualizations are serviceable, they lack creative freedom. Changing fonts, color palettes, or themes demands manual ggplot2 scripting. This deters quick stylistic adjustments and can hinder graphical consistency across reports and publications.

These limitations, though numerous, do not diminish DiaThor's value. Rather, they highlight the realities of computational ecology: every automated tool simplifies complexity at the cost of flexibility. Understanding these constraints empowers researchers to use DiaThor more intelligently – complementing it with manual validation, data visualization enhancements, and independent cross-checking. With thoughtful handling, DiaThor remains a cornerstone in the evolving landscape of diatom-based water quality assessment.

LIMITATIONS AT A GLANCE...

GRAPHICAL / VISUALIZATION

- LIMITED GRAPHICAL CUSTOMIZATION (COLORS, FONTS, THEMES REQUIRE MANUAL TWEAKING, E.G., GGPLOT2)
- LIMITED PLOT FLEXIBILITY (DEFAULT PLOTS ARE STATIC AND MINIMALLY CUSTOMIZABLE)
- STATIC PLOTS (NON-INTERACTIVE PLOTS LIMIT EXPLORATION)
- PLOT LABEL OVERLAPS (LABELS CAN OVERLAP IN NETWORK/GRAPH VISUALIZATIONS)



TAXONOMIC / DATABASE

- SPECIES NAME MISMATCH (SOME TAXA ARE UNRECOGNIZED IN REFERENCE DB → MISSING INDEX VALUES)
- TAXONOMIC UPDATES NEEDED (DATABASE NEEDS CONSTANT UPDATING FOR NEW TAXONOMY)
- DEPENDENCE ON REFERENCE DB (RESULTS RELY HEAVILY ON COMPLETENESS/ACCURACY OF DB)

INTEGRATION / COMPATIBILITY

- DIFFICULTY INTEGRATING WITH OTHER SOFTWARE (EXPORTING RESULTS TO QGIS OR EXCEL NEEDS EXTRA STEPS)
- DEPENDENCY ON OTHER PACKAGES (VERSION CONFLICTS MAY OCCUR WITH IGRAPH, GGPLOT2, ETC.)
- NOT DESIGNED FOR TIME-SERIES VISUALIZATION (TEMPORAL TRENDS REQUIRE CUSTOM HANDLING)
- LIMITED SUPPORT FOR MULTI-SITE COMPARISONS (COMPARISONS ACROSS SITES REQUIRE EXTRA DATA WRANGLING)
- NO BUILT-IN INTERACTIVE DASHBOARDS (INTERACTIVITY REQUIRES SHINY OR OTHER TOOLS)

USER / LEARNING

- LEARNING CURVE (REQUIRES KNOWLEDGE OF R AND DATA HANDLING)
- MINIMAL DOCUMENTATION FOR ADVANCED FEATURES (SOME ADVANCED FUNCTIONS ARE POORLY DOCUMENTED)
- NO MULTI-LANGUAGE SUPPORT (ALL FUNCTIONS AND ERROR MESSAGES ARE IN ENGLISH)

DATA / PROCESSING

- DATA PREPROCESSING REQUIRED (INPUT CSV MUST FOLLOW STRICT FORMATTING RULES)
- NO AUTOMATIC DATA CLEANING (USER MUST REMOVE DUPLICATES OR FIX MINOR ERRORS MANUALLY)
- NO REAL-TIME DATA VALIDATION (MISFORMATTED/MISSING DATA FLAGGED ONLY AT PROCESSING)
- LIMITED ERROR FEEDBACK (ERRORS CAN BE CRYPTIC OR SILENT)
- MEMORY AND PERFORMANCE (LARGE DATASETS MAY SLOW DOWN OR CRASH R)

FUNCTIONAL / ANALYTICAL

- INDEX COVERAGE BIAS (ONLY SOME ECOLOGICAL INDICES ARE INCLUDED; SOME MISSING)
- NO BUILT-IN CONFIDENCE OR SENSITIVITY ANALYSIS (UNCERTAINTY OF INDICES NOT PROVIDED)
- LIMITED FUNCTIONAL TRAIT INTEGRATION (ADVANCED TRAIT ANALYSES REQUIRE MANUAL WORK)
- HANDLING OF RARE SPECIES (RARE TAXA MAY BE IGNORED OR MISREPRESENTED)
- LIMITED CUSTOM STATISTICS (ADDING NEW INDICES NEEDS PROGRAMMING)



APPENDIX

This section provides supplementary information to support the use of DiaThor in diatom-based water quality analysis. Example input files (Sample_Abundance.csv, Environmental_Parameters.csv, and Site_Metadata.csv) are provided to illustrate the required data format. Basic R commands such as `library(DiaThor)`, `diaThorAll()`, and sample files that have been used here for illustrations to help readers execute and export results efficiently.

All analyses were performed in RStudio (v4.3) using DiaThor (v2.0.0). For further reference, visit <https://cran.r-project.org/package=DiaThor> or <https://diatoms.org>

This concise section ensures readers can replicate procedures, interpret outputs, and troubleshoot errors independently. You can get access of the sample files and results file, used in the book using this URL <https://github.com/Ar108-bot>

REFERENCES

- Srivastava, A., Singhal, S., Yadav, A., Zehra, R., & Verma, J. (2026). *Harnessing RStudio package for ecological insights: Monitoring Diatoms with DiaThor*. *Computational Ecology and Software* (ISSN 2220-721X). (n.d.). Retrieved September 13, 2025, from [http://www.iaees.org/publications/journals/ces/articles/2026-16\(1\)/2026-16\(1\).asp](http://www.iaees.org/publications/journals/ces/articles/2026-16(1)/2026-16(1).asp)
- Algae of India, Volume 4: A Checklist of Indian Diatoms*. (n.d.). Retrieved June 23, 2025, from <https://www.nhbs.com/algae-of-india-volume-4-book>
- A new freshwater gomphonemoid diatom genus from India, with the description of a new species from the Eastern Ghats: Phycologia: Vol 62, No 5—Get Access*. (n.d.). Retrieved June 23, 2025, from <https://www.tandfonline.com/doi/full/10.1080/00318884.2023.2268381>
- Karthick, B., Yogeshwaran, M., & Kociolek, J. P. (2023). A new freshwater gomphonemoid diatom genus from India, with the description of a new species from the Eastern Ghats. *Phycologia*, 62(5), 499–511. <https://doi.org/10.1080/00318884.2023.2268381>
- Diatoms of North America*. <https://diatoms.org/what-are-diatoms>
- AGRBIO. (n.d.). Forensic Database Generation of Diatom Species with Respect to Seasonal Variations in Ganga and Yamuna. *Scilight Press*. Retrieved June 23, 2025, from <https://agribiop.com/forensic-database-generation-of-diatom-species-with-respect-to-seasonal-variations-in-ganga-and-yamuna/>
- Caeiro, D., Aich, S., Raiborde, M. D., Katare, N., Kumar, R., & Singh, P. (2024). *Diatomological Study on Water Bodies in Bengaluru City, India*. 17(3). https://www.rfpl.co.in/view_abstract.php?id=5&art_id=15029
- Encyclopedia of Diatoms Found in the USA* · iNaturalist. (n.d.). iNaturalist. Retrieved June 23, 2025, from <https://www.inaturalist.org/guides/19954>
- Search Diatom Image Database*. (n.d.). Retrieved June 23, 2025, from <https://websites.rbge.org.uk/ADIAC/db/adiacdb.htm>
- AlgaeBase: Listing the World's Algae*. (n.d.). Retrieved June 23, 2025, from <https://www.algaebase.org/>
- Networking*. (n.d.). *International Society for Diatom Research*. Retrieved June 23, 2025, from <https://isdr.org/networking/>
- Applicability and efficacy of diatom indices in water quality evaluation of the Chambal River in Central India | Environmental Science and Pollution Research*. (n.d.). Retrieved June 23, 2025, from <https://link.springer.com/article/10.1007/s11356-017-0166-0>
- Bhatt, J. P., Manish, K., & Pandit, M. K. (2025). Diatom species assemblages and their environmental drivers in the major biogeographic provinces of India. *Proceedings of the Indian National Science Academy*. <https://doi.org/10.1007/s43538-025-00412-3>
- Diversity and ecology of freshwater diatoms as pollution indicators from the freshwater Ponds of Kanyakumari district, Tamilnadu—ScienceDirect*. (n.d.). Retrieved June 23, 2025, from <https://www.sciencedirect.com/science/article/pii/S277242712200119X>
- Full article: The diatom genus Luticola D.G. Mann (Bacillariophyceae) in the Western Ghats of India and its biogeography*. (n.d.). Retrieved June 23, 2025, from <https://www.tandfonline.com/doi/full/10.1080/09670262.2020.1783460>
- Kale, A., & Karthick, B. (2015). The diatoms. *Resonance*, 20(10), 919–930. <https://doi.org/10.1007/s12045-015-0256-6>
- Mohan, R., Shanvas, S., Thamban, M., & Sudhakar, M. (2006). Spatial distribution of diatoms in surface sediments from the Indian sector of Southern Ocean. *Current Science*, 91(11), 1495–1502.
- Nautiyal, P., Nautiyal, R., Kala, K., & Verma, J. (2004). Taxonomic richness in the diatom flora of Himalayan streams (Garhwal, India). *Diatom*, 20, 123–132. https://doi.org/10.11464/diatom1985.20.0_123
- Pareek, R., Singh, G. P., & Singh, R. (n.d.). *Some fresh water diatoms of Galta kund, Jaipur, India*.
- Srivastava, P., Verma, J., Grover, S., & Sardar, A. (2016). *ON THE IMPORTANCE OF DIATOMS AS ECOLOGICAL INDICATORS IN RIVER ECOSYSTEMS: A REVIEW*.

- Subrahmanyam, R. (1946). A systematic account of the marine plankton diatoms of the Madras coast. *Proceedings / Indian Academy of Sciences*, 24(4), 85–197. <https://doi.org/10.1007/BF03049673>
- Zoological Survey of India. Fauna list. (n.d.). Retrieved June 23, 2025, from <https://zsi.gov.in/impo-link/cn?v=7>
- India Biodiversity Portal. (n.d.). India Biodiversity Portal. Retrieved June 23, 2025, from <https://indiabiodiversity.org/>
- Ratheesh Kumar, R., Vineetha, G., & Monolisha, S. (2024). *Plankton taxonomy* (No. 41; Issue 41, pp. 155–174). ICAR- Central Marine Fisheries Research Institute. <http://eprints.cmfri.org.in/18443/>
- Diatom identification, databases and resources.* (n.d.). Irish Diatoms - Diatom Cleaning - Diatom Mounting - Micromanipulators. Retrieved June 23, 2025, from <http://www.diatomsireland.com/diatom-web-resources/>
- Leblanc, K., Aristegui, J., Armand, L., Assmy, P., Beker, B., Bode, A., Breton, E., Cornet, V., Gibson, J., Gosselin, M.-P., Kopczynska, E., Marshall, H., Peloquin, J., Piontkovski, S., Poulton, A. J., Quéguiner, B., Schiebel, R., Shipe, R., Stefels, J., ... Yallop, M. (2012). *A global diatom database – abundance, biovolume and biomass in the world ocean*. *Biological oceanography*. <https://doi.org/10.5194/essdd-5-147-2012>
- CMR Preview. (n.d.). Retrieved June 23, 2025, from <https://cmr.earthdata.nasa.gov/search/concepts/C1214597840-SCIOPS.html>
- Brylka, K., Ashworth, M. P., Alverson, A. J., & Conley, D. J. (2024). The Cretaceous Diatom Database: A tool for investigating early diatom evolution. *Journal of Phycology*, 60(5), 1090–1104. <https://doi.org/10.1111/jpy.13499>
- CAS Catalogue of Diatom Names. (n.d.). Retrieved June 23, 2025, from <https://researcharchive.calacademy.org/research/diatoms/names/index.asp>
- Diatoms – Algal Image Database.* (n.d.). Retrieved June 23, 2025, from <https://iaps.org.in/lalit/diatoms-2/>
- Diatom Image Database.* Communications, F. I. U.-D. (n.d.). Home. Retrieved June 23, 2025, from <https://fcelter.fiu.edu/>
- Armand, L., De Deckker, P., & Pichon, J.-J. (n.d.). *Antarctic Diatom Database (based on 115 surface samples)*. Research Data Australia. Retrieved June 23, 2025, from <https://researchdata.edu.au/antarctic-diatom-database-surface-samples/678697>
- Acosta-Arreola, J., Domínguez-Hüttinger, E., Aguirre, P., González, N., & Meave, J. A. (2023). Predicting dynamic trajectories of a protected plant community under contrasting conservation regimes: Insights from data-based modelling. *Ecological Modelling*, 484, 110449. <https://doi.org/10.1016/j.ecolmodel.2023.110449>
- AlgaeBase: Listing the World's Algae.* (n.d.). Retrieved June 23, 2025, from <https://www.algaebase.org/>
- Álvarez-Blanco, I., Blanco, S., Cejudo-Figueiras, C., & Bécáres, E. (2013). The Duero Diatom Index (DDI) for river water quality assessment in NW Spain: Design and validation. *Environmental Monitoring and Assessment*, 185(1), 969–981. <https://doi.org/10.1007/s10661-012-2607-z>
- Armand, L., De Deckker, P., & Pichon, J.-J. (n.d.). *Antarctic Diatom Database (based on 115 surface samples)*. Research Data Australia. Retrieved June 23, 2025, from <https://researchdata.edu.au/antarctic-diatom-database-surface-samples/678697>
- Assessing pollution of aquatic environments with diatoms' DNA metabarcoding: Experience and developments from France Water Framework Directive networks.* (n.d.). Retrieved May 14, 2025, from <https://mbmg.pensoft.net/article/39646>
- Automated Diatom Classification (Part A): Handcrafted Feature Approaches.* (n.d.). Retrieved May 14, 2025, from <https://www.mdpi.com/2076-3417/7/8/753>
- Beebe, N. H. F., & Rm, E. (n.d.). *A Complete Bibliography of Publications in Ecological Modelling (2020–2029)*.
- Benoiston, A.-S., Ibarbalz, F. M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., & Bowler, C. (2017). The evolution of diatoms and their biogeochemical functions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1728), 20160397. <https://doi.org/10.1098/rstb.2016.0397>
- BG - Multi-nutrient, multi-group model of present and future oceanic phytoplankton communities. (n.d.). Retrieved June 23, 2025, from <https://bg.copernicus.org/articles/3/585/2006/>

- Blanco, S. (n.d.). *Álvarez-Blanco I, Blanco S, Cejudo-Figueiras C, Bécades E 2012. The Duero Diatom Index (DDI) for river water quality assessment in NW Spain: Design and validation. Environ Monit Assess 185: 969-981.* <https://doi.org/10.1007/S10661-012-2607-Z>
- Bohan, D. A., Gravel, D., Tamaddon-Nezhad, A., Vacher, C., & Robin, S. (Eds.). (2020). *A Next-Generation of Biomonitoring to Detect Global Ecosystem Change*. Frontiers Media SA. <https://doi.org/10.3389/978-2-88966-027-8>
- Borja, C., & Blanco, S. (n.d.). *Blanco S, Álvarez-Blanco I, Cejudo-Figueiras C, Recio JM, Borja C, Bécades E, Díaz F, Cámara R 2013. The diatom flora in temporary ponds of Doñana National Park (southwest Spain): Five new taxa. Nordic Journal of Botany 31: 489-499.* <https://doi.org/10.1111/J.1756-1051.2013.01691.X>
- Brylka, K., Ashworth, M. P., Alverson, A. J., & Conley, D. J. (2024). The Cretaceous Diatom Database: A tool for investigating early diatom evolution. *Journal of Phycology*, 60(5), 1090–1104. <https://doi.org/10.1111/jpy.13499>
- Calculating autoecological data (optima and tolerance range) for multiple species with the 'optimos.prime' R package—Sathicq—2020—Austral Ecology—Wiley Online Library.* (n.d.). Retrieved May 15, 2025, from <https://onlinelibrary.wiley.com/doi/10.1111/aec.12868>
- CAS Catalogue of Diatom Names. (n.d.). Retrieved June 23, 2025, from <https://researcharchive.calacademy.org/research/diatoms/names/index.asp>
- ChatGPT - Ecological Insights with DiaThor. (n.d.). ChatGPT. Retrieved May 12, 2025, from <https://chatgpt.com/share/6821b5c1-bad8-8000-b0b4-a0c3dedc3570>
- CMR Preview. (n.d.). Retrieved June 23, 2025, from <https://cmr.earthdata.nasa.gov/search/concepts/C1214597840-SCIOPS.html>
- Communications, F. I. U.-D. (n.d.). *Home*. Retrieved June 23, 2025, from <https://fcelter.fiu.edu/>
- Community Characteristics Analysis of Eukaryotic Microplankton via ITS Gene Metabarcoding Based on Environmental DNA in Lower Reaches of Qiantang River, China.* (n.d.). Retrieved May 14, 2025, from [https://www.scirp.org/\(S\(d0vxmy1blcw245otj11h3a\)\)/journal/paperinformation?paperid=108225](https://www.scirp.org/(S(d0vxmy1blcw245otj11h3a))/journal/paperinformation?paperid=108225)
- Coste, M. & Ricard, M. , 1983 “1982”: *AlgaeBase*. (n.d.). Retrieved May 14, 2025, from https://www.algaebase.org/search/bibliography/detail/?biblio_id=59993
- Cryptogamie. Algologie.* (n.d.).
- DiaThor: R package for computing diatom metrics and biotic indices* | bioRxiv. (n.d.). Retrieved May 15, 2025, from <https://www.biorxiv.org/content/10.1101/2021.03.31.437895v2>
- Diatom eDNA metabarcoding and morphological methods for bioassessment of karstic river | Request PDF. (n.d.). ResearchGate. <https://doi.org/10.1016/j.scitotenv.2022.154536>
- Diatom identification, databases and resources.* (n.d.). Irish Diatoms - Diatom Cleaning - Diatom Mounting - Micromanipulators. Retrieved June 23, 2025, from <http://www.diatomsireland.com/diatom-web-resources/>
- Diatom—An overview* | ScienceDirect Topics. (n.d.). Retrieved June 23, 2025, from <https://www.sciencedirect.com/topics/earth-and-planetary-sciences/diatom>
- Diatoms – Algal Image Database.* (n.d.). Retrieved June 23, 2025, from <https://iaps.org.in/lalit/diatoms-2/>
- Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for Community action in the field of water policy—European Environment Agency.* (n.d.). Retrieved May 14, 2025, from <https://www.eea.europa.eu/policy-documents/directive-2000-60-ec-of>
- Encyclopedia of Diatoms Found in the USA · iNaturalist.* (n.d.-a). iNaturalist. Retrieved June 23, 2025, from <https://www.inaturalist.org/guides/19954>
- Encyclopedia of Diatoms Found in the USA · iNaturalist.* (n.d.-b). iNaturalist. Retrieved June 23, 2025, from <https://www.inaturalist.org/guides/19954>
- Gelis, M. M. N., Sathicq, M. B., & Cochero, J. (2024). *diathor: Calculate Ecological Information and Diatom Based Indices (Version 0.1.5)* [Computer software]. <https://cran.r-project.org/web/packages/diathor/index.html>

Impact of extreme drought on diatom traits and species composition in temperate lowland streams | *Hydrobiologia*. (n.d.). Retrieved May 15, 2025, from <https://link.springer.com/article/10.1007/s10750-024-05710-3>

Jjallaire · *GitHub*. (n.d.). Retrieved May 16, 2025, from <https://github.com/jjallaire>

Joaquin Cochero (0000-0003-3957-6819)—ORCID. (n.d.). Retrieved May 15, 2025, from <https://orcid.org/0000-0003-3957-6819>

Kang, W., Anslan, S., Börner, N., Schwarz, A., Schmidt, R., Künzel, S., Rioual, P., Echeverría-Galindo, P., Vences, M., Wang, J., & Schwalb, A. (2021). Diatom metabarcoding and microscopic analyses from sediment samples at Lake Nam Co, Tibet: The effect of sample-size and bioinformatics on the identified communities. *Ecological Indicators*, 121, 107070. <https://doi.org/10.1016/j.ecolind.2020.107070>

Keck, F. (2025). *Fkeck/diatbarcode* [R]. <https://github.com/fkeck/diatbarcode> (Original work published 2019)

Keck, F., Vasselon, V., Tapolczai, K., Rimet, F., & Bouchez, A. (2017). Freshwater biomonitoring in the Information Age. *Frontiers in Ecology and the Environment*, 15(5), 266–274. <https://doi.org/10.1002/fec.1490>

Kelly, M. G., & Whitton, B. A. (1995). The Trophic Diatom Index: A new index for monitoring eutrophication in rivers. *Journal of Applied Phycology*, 7(4), 433–444. <https://doi.org/10.1007/BF00003802>

Leblanc, K., Aristegui, J., Armand, L., Assmy, P., Beker, B., Bode, A., Breton, E., Cornet, V., Gibson, J., Gosselin, M.-P., Kopczynska, E., Marshall, H., Pelloquin, J., Piontkovski, S., Poulton, A. J., Quéguiner, B., Schiebel, R., Shipe, R., Stefels, J., ... Yallop, M. (2012). *A global diatom database – abundance, biovolume and biomass in the world ocean*. Biological oceanography. <https://doi.org/10.5194/essdd-5-147-2012>

Lecoite, C., Coste, M., & Prygiel, J. (1993a). “Omnidia”: Software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, 269(1), 509–513. <https://doi.org/10.1007/BF00028048>

Lecoite, C., Coste, M., & Prygiel, J. (1993b). ?Omnidia?: Software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, 269–270(1), 509–513. <https://doi.org/10.1007/BF00028048>

Litchman, E., Klausmeier, C. A., Miller, J. R., Schofield, O. M., & Falkowski, P. G. (2006). Multi-nutrient, multi-group model of present and future oceanic phytoplankton communities. *Biogeosciences*, 3(4), 585–606. <https://doi.org/10.5194/bg-3-585-2006>

Liu, B., Chen, S., Liu, H., & Guan, Y. (2020). Modeling cyanobacteria biomass by surface sediment diatoms in lakes: Problems and suggestions. *Ecological Modelling*, 430, 109056. <https://doi.org/10.1016/j.ecolmodel.2020.109056>

María Belén Sathicq (0000-0002-3534-8950)—ORCID. (n.d.). Retrieved May 15, 2025, from <https://orcid.org/0000-0002-3534-8950>

Maria Mercedes Nicolosi Gelis (0000-0001-6324-7930). (n.d.). ORCID. Retrieved May 15, 2025, from <https://orcid.org/0000-0001-6324-7930>

Networking. (n.d.). *International Society for Diatom Research*. Retrieved June 23, 2025, from <https://isdr.org/networking/>

Nicolosi Gelis, M. M., Cochero, J., Donadelli, J., & Gómez, N. (2020a). Exploring the use of nuclear alterations, motility and ecological guilds in epipelagic diatoms as biomonitoring tools for water quality improvement in urban impacted lowland streams. *Ecological Indicators*, 110, 105951. <https://doi.org/10.1016/j.ecolind.2019.105951>

Nicolosi Gelis, M. M., Cochero, J., Donadelli, J., & Gómez, N. (2020b). Exploring the use of nuclear alterations, motility and ecological guilds in epipelagic diatoms as biomonitoring tools for water quality improvement in urban impacted lowland streams. *Ecological Indicators*, 110, 105951. <https://doi.org/10.1016/j.ecolind.2019.105951>

Nicolosi Gelis, M. M., & Sathicq, M. B. (2020a). *diathor: Calculate Ecological Information and Diatom Based Indices* (p. 0.1.5) [Dataset]. <https://doi.org/10.32614/CRAN.package.diathor>

Nicolosi Gelis, M. M., & Sathicq, M. B. (2020b). *diathor: Calculate Ecological Information and Diatom Based Indices* (p. 0.1.5) [Dataset]. <https://doi.org/10.32614/CRAN.package.diathor>

- Nicolosi Gelis, M. M., & Sathicq, M. B. (2020c). *diathor: Calculate Ecological Information and Diatom Based Indices* (p. 0.1.5) [Dataset]. <https://doi.org/10.32614/CRAN.package.diathor>
- Nicolosi Gelis, M. M., Sathicq, M. B., Jupke, J., & Cocherio, J. (2022a). DiaThor: R package for computing diatom metrics and biotic indices. *Ecological Modelling*, 465, 109859. <https://doi.org/10.1016/j.ecolmodel.2021.109859>
- Nicolosi Gelis, M. M., Sathicq, M. B., Jupke, J., & Cocherio, J. (2022b). DiaThor: R package for computing diatom metrics and biotic indices. *Ecological Modelling*, 465, 109859. <https://doi.org/10.1016/j.ecolmodel.2021.109859>
- Nicolosi Gelis, M. M., Sathicq, M. B., Jupke, J., & Cocherio, J. (2022c). DiaThor: R package for computing diatom metrics and biotic indices. *Ecological Modelling*, 465, 109859. <https://doi.org/10.1016/j.ecolmodel.2021.109859>
- Nicolosi Gelis, M.M., Sathicq, M.B., Jupke, J., & Cocherio, J. (2021). - Google Search. (n.d.). Retrieved May 14, 2025, from https://www.google.com/search?q=Nicolosi+Gelis%2C+M.M.%2C+Sathicq%2C+M.B.%2C+Jupke%2C+J.%2C+%26+Cocherio%2C+J.%282021%29.&scas_esv=1bbf12dcdfda0a2c&sxsrf=AHTn8zqQu13Ce-butu2ALkWW5-NSOo9Emg%3A1747231072124&ei=YKEkaM6kB5CHvr0PzL6ngAs&ved=0ahUKEwjO1JP7jqONAxWOG68BHUzfCbAQ4dUDCBA&uact=5&oq=Nicolosi+Gelis%2C+M.M.%2C+Sathicq%2C+M.B.%2C+Jupke%2C+J.%2C+%26+Cocherio%2C+J.%282021%29.&gs_lp=Egxnd3Mtd2l6LXNlcniA1RU5pY29sb3NpIEdlbG1zLCBNLk0uLCBTyXRoaWNxLCBNLkluLCBKdXBrZSwgSi4sICYgQ29jaGVybywgSi4gKDIwMjEpLkgAUABYAHAAcACOAQCYAOCgAOCgAOC4AOPIAOD4AOL4AOQYAgCgAgCYAwCSBwCgBwCvBwC4BwA&scient=gws-wiz-serp
- Oeding, S., & Taffs, K. H. (2017). Developing a regional diatom index for assessment and monitoring of freshwater streams in sub-tropical Australia. *Ecological Indicators*, 80, 135–146. <https://doi.org/10.1016/j.ecolind.2017.05.009>
- Overview of Biological and Hydromorphological Assessment Methods in the Danube River Basin. (n.d.).
- Pandey, L. Kr., Ojha, K. Kr., Singh, P. K., Singh, C. S., Dwivedi, S., & Bergey, E. A. (2016). Diatoms image database of India (DIDI): A research tool. *Environmental Technology & Innovation*, 5, 148–160. <https://doi.org/10.1016/j.eti.2016.02.001>
- (PDF) DiatomNet: An automatic Diatom genus identification system through microscopic images and Deep Learning. (2025, March 29). ResearchGate. <https://doi.org/10.1101/2025.02.10.635050>
- (PDF) Successive accumulation of biotic assemblages at a fine spatial scale along glacier-fed waters. (2024). ResearchGate. <https://doi.org/10.1016/j.isci.2024.109476>
- (PDF) Using diatom life-forms and ecological guilds to assess organic pollution and trophic level in rivers: A case study of rivers in south-eastern France. (n.d.). Retrieved May 14, 2025, from https://www.researchgate.net/publication/227310241_Using_diatom_life-forms_and_ecological_guilds_to_assess_organic_pollution_and_trophic_level_in_rivers_a_case_study_of_rivers_in_south-eastern_France
- Posit. (n.d.-a). Posit. Retrieved May 16, 2025, from <https://www.posit.co/>
- Posit. (n.d.-b). Posit. Retrieved May 16, 2025, from <https://www.posit.co/>
- Pu, S., Zhang, F., Shu, Y., & Fu, W. (2023). Microscopic image recognition of diatoms based on deep learning. *Journal of Phycology*, 59(6), 1166–1178. <https://doi.org/10.1111/jpy.13390>
- Rimet, F. (n.d.). Larras F., Bouchez A., Rimet F. & Montuelle B. 2012. Using Bioassays and Species Sensitivity Distributions to Assess Herbicide Toxicity towards Benthic Diatoms. *PLoS ONE* 7 (8). <https://doi.org/10.1371/JOURNAL.PONE.0044458>
- Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A., & Bouchez, A. (2016). R-Syst::diatom: An open-access and curated barcode database for diatoms and freshwater monitoring. *Database*, 2016, baw016. <https://doi.org/10.1093/database/baw016>
- Search Diatom Image Database. (n.d.). Retrieved June 23, 2025, from <https://websites.rbge.org.uk/ADIAC/db/adiacdb.htm>

- Selvaratnam, N. (2023). *How to Run Essential Analyses in R Studio: For First Year Undergraduate Students with No Background in Programming* (Eagan Journal of Contemporary Research, Ed.; First Edition). Deep Haven Counselling (Pvt) Ltd (10.61495). <https://doi.org/10.61495/ejcr.0001.23>
- Spaulding, S. A., Potapova, M. G., Bishop, I. W., Lee, S. S., Gasperak, T. S., Jovanoska, E., Furey, P. C., & Edlund, M. B. (2021a). *Diatoms.org*: Supporting taxonomists, connecting communities. *Diatom Research*, 36(4), 291–304. <https://doi.org/10.1080/0269249X.2021.2006790>
- Spaulding, S. A., Potapova, M. G., Bishop, I. W., Lee, S. S., Gasperak, T. S., Jovanoska, E., Furey, P. C., & Edlund, M. B. (2021b). *Diatoms.org*: Supporting taxonomists, connecting communities. *Diatom Research*, 36(4), 291–304. <https://doi.org/10.1080/0269249X.2021.2006790>
- Spaulding, S. A., Potapova, M. G., Bishop, I. W., Lee, S. S., Gasperak, T. S., Jovanoska, E., Furey, P. C., & Edlund, M. B. (2021c). *Diatoms.org*: Supporting taxonomists, connecting communities. *Diatom Research*, 36(4), 291–304. <https://doi.org/10.1080/0269249X.2021.2006790>
- Spaulding, S. A., Potapova, Marina G., Bishop, Ian W., Lee, Sylvia S., Gasperak, Tim S., Jovanoska, Elena, Furey, Paula C., & Edlund, M. B. (2021). *Diatoms.org*: Supporting taxonomists, connecting communities. *Diatom Research*, 36(4), 291–304. <https://doi.org/10.1080/0269249X.2021.2006790>
- Tapolczai, K., Chonova, T., Fidlerová, D., Makovinská, J., Mora, D., Weigand, A., & Zimmermann, J. (2024a). Molecular metrics to monitor ecological status of large rivers: Implementation of diatom DNA metabarcoding in the Joint Danube Survey 4. *Ecological Indicators*, 160, 111883. <https://doi.org/10.1016/j.ecolind.2024.111883>
- Tapolczai, K., Chonova, T., Fidlerová, D., Makovinská, J., Mora, D., Weigand, A., & Zimmermann, J. (2024b). Molecular metrics to monitor ecological status of large rivers: Implementation of diatom DNA metabarcoding in the Joint Danube Survey 4. *Ecological Indicators*, 160, 111883. <https://doi.org/10.1016/j.ecolind.2024.111883>
- Teittinen, A., Taka, M., Ruth, O., & Soininen, J. (2015a). Variation in stream diatom communities in relation to water quality and catchment variables in a boreal, urbanized region. *Science of The Total Environment*, 530–531, 279–289. <https://doi.org/10.1016/j.scitotenv.2015.05.101>
- Teittinen, A., Taka, M., Ruth, O., & Soininen, J. (2015b). Variation in stream diatom communities in relation to water quality and catchment variables in a boreal, urbanized region. *Science of The Total Environment*, 530–531, 279–289. <https://doi.org/10.1016/j.scitotenv.2015.05.101>
- Tree of Science with Scopus: A Shiny Application | Issues in Science and Technology Librarianship*. (n.d.). Retrieved May 14, 2025, from <https://journals.library.ualberta.ca/istl/index.php/istl/article/view/2698>
- “UDE DIATOMS in the Wild 2024”: A new image dataset of freshwater diatoms for training deep learning models | Request PDF. (2024). *ResearchGate*. <https://doi.org/10.1093/gigascience/giae087>
- Valentin, V., Frédéric, R., Isabelle, D., Olivier, M., Yorick, R., & Agnès, B. (2019). Assessing pollution of aquatic environments with diatoms' DNA metabarcoding: Experience and developments from France Water Framework Directive networks. *Metabarcoding and Metagenomics*, 3, e39646. <https://doi.org/10.3897/mbmg.3.39646>
- Villar, P., Casado, J., Fernández, D., Sánchez, P., & Tabik, S. (2025). *DiatomNet: An automatic Diatom genus identification system through microscopic images and Deep Learning* (p. 2025.02.10.635050). *bioRxiv*. <https://doi.org/10.1101/2025.02.10.635050>
- Wood, R. J., Mitrovic, S. M., Lim, R. P., Warne, M. St. J., Dunlop, J., & Kefford, B. J. (2019). Benthic diatoms as indicators of herbicide toxicity in rivers – A new SPEcies At Risk (SPEARherbicides) index. *Ecological Indicators*, 99, 203–213. <https://doi.org/10.1016/j.ecolind.2018.12.035>
- Yu, W., Xiang, Q., Hu, Y., Du, Y., Kang, X., Zheng, D., Shi, H., Xu, Q., Li, Z., Niu, Y., Liu, C., & Zhao, J. (2022). An improved automated diatom detection method based on YOLOv5 framework and its preliminary study for taxonomy recognition in the forensic diatom test. *Frontiers in Microbiology*, 13. <https://doi.org/10.3389/fmicb.2022.963059>
- Dam, H., Mertens, A., & Sinkeldam, J. (1994). A coded checklist and ecological indicator values of freshwater diatoms from The Netherlands. *Netherlands Journal of Aquatic Ecology*, 28(1), 117–133. <https://doi.org/10.1007/bf02334251>

- Lecointe, C., Coste, M., & Prygiel, J. (1993). "Omnia": Software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, 269(1), 509–513. <https://doi.org/10.1007/BF00028048>
- Mertens, A., van der Wal, J., Verweij, G., Pex, B., van Dulmen, A., & van Dam, H. (2025). A revised list of diatom ecological indicator values in The Netherlands. *Ecological Indicators*, 172, 113219. <https://doi.org/10.1016/j.ecolind.2025.113219>
- (PDF) A Coded Check List and Ecological Indicator Values of Freshwater Diatoms From the Netherlands. (2025). ResearchGate. <https://doi.org/10.1007/BF02334251>
- (PDF) Comparison of diatom-based indices of water quality of mid-continent (USA) great rivers. (n.d.). ResearchGate. Retrieved July 7, 2025, from https://www.researchgate.net/publication/286751874_Comparison_of_diatom-based_indices_of_water_quality_of_mid-continent_USA_great_rivers
- Search results | Taylor & Francis Online. (n.d.). Retrieved July 7, 2025, from <https://www.tandfonline.com/action/doSearch?AllField=%09SI%C3%A1de%C4%8Dek%27s+Saprobic&Index&SeriesKey=tinw19&startPage=&pageSize=10>
- Sládeček, V., & Sládečková, A. (1998). Revision of polysaprobic indicators. *SIL Proceedings, 1922-2010*, 26(3), 1277–1280. <https://doi.org/10.1080/03680770.1995.11900928>
- Spaulding, S. A., Potapova, M. G., Bishop, I. W., Lee, S. S., Gasperak, T. S., Jovanoska, E., Furey, P. C., & Edlund, M. B. (2021). Diatoms.org: Supporting taxonomists, connecting communities. *Diatom Research*, 36(4), 291–304. <https://doi.org/10.1080/0269249x.2021.2006790>
- Szczepocka, E., & Szulc, B. (2009). The use of benthic diatoms in estimating water quality of variously polluted rivers. *Oceanological and Hydrobiological Studies*, 38(1), 17–26. <https://doi.org/10.2478/v10009-009-0012-x>
- The measures of saprobity: With 6 figures and 5 tables in the text: *SIL Proceedings, 1922-2010: Vol 17, No 1*. (n.d.). Retrieved July 7, 2025, from <https://www.tandfonline.com/doi/abs/10.1080/03680770.1968.11895887>
- Trophic and saprobic diatom-based indices are not indicating the same ecological status in a shallow lake, Türkiye | Request PDF. (n.d.). ResearchGate. <https://doi.org/10.1007/s42974-024-00199-1>
- Wood, R. J., Mitrovic, S. M., Lim, R. P., Warne, M. St. J., Dunlop, J., & Kefford, B. J. (2019). Benthic diatoms as indicators of herbicide toxicity in rivers – A new SPEcies At Risk (SPEARherbicides) index. *Ecological Indicators*, 99, 203–213. <https://doi.org/10.1016/j.ecolind.2018.12.035>
- A new freshwater gomphonemoid diatom genus from India, with the description of a new species from the Eastern Ghats: Phycologia: Vol 62 , No 5—Get Access.* (n.d.). Retrieved June 23, 2025, from <https://www.tandfonline.com/doi/full/10.1080/00318884.2023.2268381>
- Acosta-Arreola, J., Domínguez-Hüttinger, E., Aguirre, P., González, N., & Meave, J. A. (2023). Predicting dynamic trajectories of a protected plant community under contrasting conservation regimes: Insights from data-based modelling. *Ecological Modelling*, 484, 110449. <https://doi.org/10.1016/j.ecolmodel.2023.110449>
- AGRBIO. (n.d.). Forensic Database Generation of Diatom Species with Respect to Seasonal Variations in Ganga and Yamuna. *SciLight Press*. Retrieved June 23, 2025, from <https://agribiop.com/forensic-database-generation-of-diatom-species-with-respect-to-seasonal-variations-in-ganga-and-yamuna/>
- Algae of India, Volume 4: A Checklist of Indian Diatoms.* (n.d.). Retrieved June 23, 2025, from <https://www.nhbs.com/algae-of-india-volume-4-book>
- AlgaeBase: Listing the World's Algae.* (n.d.). Retrieved June 23, 2025, from <https://www.algaebase.org/>
- Álvarez-Blanco, I., Blanco, S., Cejudo-Figueiras, C., & Bécares, E. (2013). The Duero Diatom Index (DDI) for river water quality assessment in NW Spain: Design and validation. *Environmental Monitoring and Assessment*, 185(1), 969–981. <https://doi.org/10.1007/s10661-012-2607-z>
- Applicability and efficacy of diatom indices in water quality evaluation of the Chambal River in Central India | Environmental Science and Pollution Research.* (n.d.). Retrieved June 23, 2025, from <https://link.springer.com/article/10.1007/s11356-017-0166-0>

- Armand, L., De Deckker, P., & Pichon, J.-J. (n.d.). *Antarctic Diatom Database (based on 115 surface samples)*. Research Data Australia. Retrieved June 23, 2025, from <https://researchdata.edu.au/antarctic-diatom-database-surface-samples/678697>
- Assessing pollution of aquatic environments with diatoms' DNA metabarcoding: Experience and developments from France Water Framework Directive networks.* (n.d.). Retrieved May 14, 2025, from <https://mbmg.pensoft.net/article/39646>
- Automated Diatom Classification (Part A): Handcrafted Feature Approaches.* (n.d.). Retrieved May 14, 2025, from <https://www.mdpi.com/2076-3417/7/8/753>
- Beebe, N. H. F., & Rm, E. (n.d.). *A Complete Bibliography of Publications in Ecological Modelling (2020–2029)*.
- Beeley, C., & Sukhdeve, S. R. (2018). *Web Application Development with R Using Shiny: Build stunning graphics and interactive data visualizations to deliver cutting-edge analytics, 3rd Edition*. Packt Publishing Ltd.
- Benoiston, A.-S., Ibarbalz, F. M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., & Bowler, C. (2017). The evolution of diatoms and their biogeochemical functions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1728), 20160397. <https://doi.org/10.1098/rstb.2016.0397>
- BG - Multi-nutrient, multi-group model of present and future oceanic phytoplankton communities. (n.d.). Retrieved June 23, 2025, from <https://bg.copernicus.org/articles/3/585/2006/>
- Bhatt, J. P., Manish, K., & Pandit, M. K. (2025). Diatom species assemblages and their environmental drivers in the major biogeographic provinces of India. *Proceedings of the Indian National Science Academy*. <https://doi.org/10.1007/s43538-025-00412-3>
- Blanco, S. (n.d.). Álvarez-Blanco I, Blanco S, Cejudo-Figueiras C, Bécares E 2012. The Duero Diatom Index (DDI) for river water quality assessment in NW Spain: Design and validation. *Environ Monit Assess* 185: 969-981. <https://doi.org/10.1007/S10661-012-2607-Z>
- Bohan, D. A., Gravel, D., Tamaddon-Nezhad, A., Vacher, C., & Robin, S. (Eds.). (2020). *A Next-Generation of Biomonitoring to Detect Global Ecosystem Change*. Frontiers Media SA. <https://doi.org/10.3389/978-2-88966-027-8>
- Borja, C., & Blanco, S. (n.d.). Blanco S, Álvarez-Blanco I, Cejudo-Figueiras C, Recio JM, Borja C, Bécares E, Díaz F, Cámara R 2013. The diatom flora in temporary ponds of Doñana National Park (southwest Spain): Five new taxa. *Nordic Journal of Botany* 31: 489-499. <https://doi.org/10.1111/J.1756-1051.2013.01691.X>
- Brylka, K., Ashworth, M. P., Alverson, A. J., & Conley, D. J. (2024). The Cretaceous Diatom Database: A tool for investigating early diatom evolution. *Journal of Phycology*, 60(5), 1090–1104. <https://doi.org/10.1111/jpy.13499>
- Caeiro, D., Aich, S., Raiborde, M. D., Katare, N., Kumar, R., & Singh, P. (2024). *Diatomological Study on Water Bodies in Bengaluru City, India*. 17(3).
- Calculating autoecological data (optima and tolerance range) for multiple species with the 'optimos.prime' R package—Sathicq—2020—Austral Ecology—Wiley Online Library.* (n.d.). Retrieved May 15, 2025, from <https://onlinelibrary.wiley.com/doi/10.1111/aec.12868>
- CAS Catalogue of Diatom Names. (n.d.). Retrieved June 23, 2025, from <https://researcharchive.calacademy.org/research/diatoms/names/index.asp>
- ChatGPT - Ecological Insights with DiaThor. (n.d.). ChatGPT. Retrieved May 12, 2025, from <https://chatgpt.com/share/6821b5c1-bad8-8000-b0b4-a0c3dedc3570>
- CMR Preview. (n.d.). Retrieved June 23, 2025, from <https://cmr.earthdata.nasa.gov/search/concepts/C1214597840-SCIOPS.html>
- Communications, F. I. U.-D. (n.d.). *Home*. Retrieved June 23, 2025, from <https://fcelter.fiu.edu/>
- Community Characteristics Analysis of Eukaryotic Microplankton via ITS Gene Metabarcoding Based on Environmental DNA in Lower Reaches of Qiantang River, China.* (n.d.). Retrieved May 14, 2025, from [https://www.scirp.org/\(S\(dit0vxmy1blcwx245otj11h3a\)\)/journal/paperinformation?paperid=108225](https://www.scirp.org/(S(dit0vxmy1blcwx245otj11h3a))/journal/paperinformation?paperid=108225)
- Coste, M. & Ricard, M. , 1983 "1982": *AlgaeBase*. (n.d.). Retrieved May 14, 2025, from https://www.algaebase.org/search/bibliography/detail/?biblio_id=59993

Cryptogamie. Algologie. (n.d.).

Dam, H., Mertens, A., & Sinkeldam, J. (1994). A coded checklist and ecological indicator values of freshwater diatoms from The Netherlands. *Netherlands Journal of Aquatic Ecology*, 28(1), 117–133. <https://doi.org/10.1007/bf02334251>

D'Costa, P. M., & Anil, A. C. (2010). Diatom community dynamics in a tropical, monsoon-influenced environment: West coast of India. *Continental Shelf Research*, 30(12), 1324–1337. <https://doi.org/10.1016/j.csr.2010.04.015>

DiaThor: R package for computing diatom metrics and biotic indices | bioRxiv. (n.d.). Retrieved May 15, 2025, from <https://www.biorxiv.org/content/10.1101/2021.03.31.437895v2>

Diatom eDNA metabarcoding and morphological methods for bioassessment of karstic river | Request PDF. (n.d.). ResearchGate. <https://doi.org/10.1016/j.scitotenv.2022.154536>

Diatom identification, databases and resources. (n.d.). Irish Diatoms - Diatom Cleaning - Diatom Mounting - Micromanipulators. Retrieved June 23, 2025, from <http://www.diatomsireland.com/diatom-web-resources/>

Diatom—An overview | ScienceDirect Topics. (n.d.). Retrieved June 23, 2025, from <https://www.sciencedirect.com/topics/earth-and-planetary-sciences/diatom>

Diatoms – Algal Image Database. (n.d.). Retrieved June 23, 2025, from <https://iaps.org.in/lalit/diatoms-2/>

Diatoms (Bacillariophyceae) from Orissa State and Neighbouring Regions, India. (n.d.). Retrieved June 23, 2025, from <https://e-algae.org/journal/view.php?doi=10.4490/algae.2006.21.4.377>

Diatoms (Bacillariophyceae) from Orissa State and Neighbouring Regions, India -ALGAE | Korea Science. (n.d.). Retrieved June 23, 2025, from <https://koreascience.kr/article/JAKO200616049040487.page>

Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for Community action in the field of water policy—European Environment Agency. (n.d.). Retrieved May 14, 2025, from <https://www.eea.europa.eu/policy-documents/directive-2000-60-ec-of>

Diversity and ecology of freshwater diatoms as pollution indicators from the freshwater Ponds of Kanyakumari district, Tamilnadu—ScienceDirect. (n.d.). Retrieved June 23, 2025, from <https://www.sciencedirect.com/science/article/pii/S277242712200119X>

Encyclopedia of Diatoms Found in the USA · iNaturalist. (n.d.-a). iNaturalist. Retrieved June 23, 2025, from <https://www.inaturalist.org/guides/19954>

Encyclopedia of Diatoms Found in the USA · iNaturalist. (n.d.-b). iNaturalist. Retrieved June 23, 2025, from <https://www.inaturalist.org/guides/19954>

Full article: The diatom genus Luticola D.G. Mann (Bacillariophyceae) in the Western Ghats of India and its biogeography. (n.d.). Retrieved June 23, 2025, from <https://www.tandfonline.com/doi/full/10.1080/09670262.2020.1783460>

Gelis, M. M. N., & Sathicq, M. B. (2020, November 26). *Calculate Ecological Information and Diatom Based Indices [R package diathor version 0.0.1]*. <https://www.semanticscholar.org/paper/Calculate-Ecological-Information-and-Diatom-Based-Gelis-Sathicq/5b02def1915ddb8afbb529646b67dcce7172718e>

Gelis, M. M. N., Sathicq, M. B., & Cochero, J. (2024). *diathor: Calculate Ecological Information and Diatom Based Indices* (Version 0.1.5) [Computer software]. <https://cran.r-project.org/web/packages/diathor/index.html>

Impact of extreme drought on diatom traits and species composition in temperate lowland streams | Hydrobiologia. (n.d.). Retrieved May 15, 2025, from <https://link.springer.com/article/10.1007/s10750-024-05710-3>

India Biodiversity Portal. (n.d.). India Biodiversity Portal. Retrieved June 23, 2025, from <https://indiabiodiversity.org/>

Jena, M., Ratha, S. K., & Adhikary, S. P. (2006). Diatoms (Bacillariophyceae) from Orissa State and Neighbouring Regions, India. *ALGAE*, 21(4), 377–392. <https://doi.org/10.4490/ALGAE.2006.21.4.377>

Jia, L., Yao, W., Jiang, Y., Li, Y., Wang, Z., Li, H., Huang, F., Li, J., Chen, T., & Zhang, H. (2022). Development of interactive biological web applications with R/Shiny. *Briefings in Bioinformatics*, 23(1), bbab415. <https://doi.org/10.1093/bib/bbab415>

Jjallaire · *GitHub*. (n.d.). Retrieved May 16, 2025, from <https://github.com/jjallaire>

Joaquin Cochero (0000-0003-3957-6819)—*ORCID*. (n.d.). Retrieved May 15, 2025, from <https://orcid.org/0000-0003-3957-6819>

Just a moment... (n.d.). Retrieved June 23, 2025, from https://www.researchgate.net/profile/Balasubramanian-Karthick/publication/233615866_The_Diatom_Genus_Gomphonema_Ehrenberg_in_India_checklist_and_description_of_three_new_species/links/00b495209da8db1ee1000000/The-Diatom-Genus-Gomphonema-Ehrenberg-in-India-checklist-and-description-of-three-new-species.pdf

Kale, A., & Karthick, B. (2015). The diatoms. *Resonance*, 20(10), 919–930. <https://doi.org/10.1007/s12045-015-0256-6>

Kang, W., Anslan, S., Börner, N., Schwarz, A., Schmidt, R., Künzel, S., Rioual, P., Echeverria-Galindo, P., Vences, M., Wang, J., & Schwalb, A. (2021). Diatom metabarcoding and microscopic analyses from sediment samples at Lake Nam Co, Tibet: The effect of sample-size and bioinformatics on the identified communities. *Ecological Indicators*, 121, 107070. <https://doi.org/10.1016/j.ecolind.2020.107070>

Karthick, B., Kociolek, J. P., Mahesh, M. K., & Ramachandra, T. V. (2011a). The diatom genus Gomphonema Ehrenberg in India: Checklist and description of three new species. *Nova Hedwigia*, 93(1–2), 211–236. <https://doi.org/10.1127/0029-5035/2011/0093-0211>

Karthick, B., Kociolek, J. P., Mahesh, M. K., & Ramachandra, T. V. (2011b). The diatom genus Gomphonema Ehrenberg in India: Checklist and description of three new species. *Nova Hedwigia*, 93(1–2), 211–236. <https://doi.org/10.1127/0029-5035/2011/0093-0211>

Karthick, B., Yogeshwaran, M., & Kociolek, J. P. (2023). A new freshwater gomphonemoid diatom genus from India, with the description of a new species from the Eastern Ghats. *Phycologia*, 62(5), 499–511. <https://doi.org/10.1080/00318884.2023.2268381>

Keck, F. (2025). *Fkeck/diatbarcode* [R]. <https://github.com/fkeck/diatbarcode> (Original work published 2019)

Keck, F., Vasselon, V., Tapolczai, K., Rimet, F., & Bouchez, A. (2017). Freshwater biomonitoring in the Information Age. *Frontiers in Ecology and the Environment*, 15(5), 266–274. <https://doi.org/10.1002/fec.1490>

Kelly, M. G., & Whitton, B. A. (1995). The Trophic Diatom Index: A new index for monitoring eutrophication in rivers. *Journal of Applied Phycology*, 7(4), 433–444. <https://doi.org/10.1007/BF00003802>

K.P., D., K., D., & B., S. (n.d.). *Chordata:Amphibia* [Dataset]. <https://doi.org/10.26515/Fauna/1/2023/Chordata:Amphibia>

Leblanc, K., Aristegui, J., Armand, L., Assmy, P., Beker, B., Bode, A., Breton, E., Cornet, V., Gibson, J., Gosselin, M.-P., Kopczynska, E., Marshall, H., Ploquin, J., Piontkowski, S., Poulton, A. J., Quéguiner, B., Schiebel, R., Shipe, R., Stefels, J., ... Yallop, M. (2012). *A global diatom database – abundance, biovolume and biomass in the world ocean*. Biological oceanography. <https://doi.org/10.5194/essdd-5-147-2012>

Lecointe, C., Coste, M., & Prygiel, J. (1993a). “Omnidia”: Software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, 269(1), 509–513. <https://doi.org/10.1007/BF00028048>

Lecointe, C., Coste, M., & Prygiel, J. (1993b). “Omnidia”: Software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, 269(1), 509–513. <https://doi.org/10.1007/BF00028048>

Lecointe, C., Coste, M., & Prygiel, J. (1993c). ?Omnidia?: Software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, 269–270(1), 509–513. <https://doi.org/10.1007/BF00028048>

Litchman, E., Klausmeier, C. A., Miller, J. R., Schofield, O. M., & Falkowski, P. G. (2006). Multi-nutrient, multi-group model of present and future oceanic phytoplankton communities. *Biogeosciences*, 3(4), 585–606. <https://doi.org/10.5194/bg-3-585-2006>

Liu, B., Chen, S., Liu, H., & Guan, Y. (2020). Modeling cyanobacteria biomass by surface sediment diatoms in lakes: Problems and suggestions. *Ecological Modelling*, 430, 109056. <https://doi.org/10.1016/j.ecolmodel.2020.109056>

María Belén Sathicq (0000-0002-3534-8950)—*ORCID*. (n.d.). Retrieved May 15, 2025, from <https://orcid.org/0000-0002-3534-8950>

- Maria Mercedes Nicolosi Gelis (0000-0001-6324-7930). (n.d.). ORCID. Retrieved May 15, 2025, from <https://orcid.org/0000-0001-6324-7930>
- Mertens, A., van der Wal, J., Verweij, G., Pex, B., van Dulmen, A., & van Dam, H. (2025). A revised list of diatom ecological indicator values in The Netherlands. *Ecological Indicators*, 172, 113219. <https://doi.org/10.1016/j.ecolind.2025.113219>
- Mohan, R., Shanvas, S., Thamban, M., & Sudhakar, M. (2006). Spatial distribution of diatoms in surface sediments from the Indian sector of Southern Ocean. *Current Science*, 91(11), 1495–1502.
- Nautiyal, P., Nautiyal, R., Kala, K., & Verma, J. (2004). Taxonomic richness in the diatom flora of Himalayan streams (Garhwal, India). *Diatom*, 20, 123–132. https://doi.org/10.11464/diatom1985.20.0_123
- Networking. (n.d.). *International Society for Diatom Research*. Retrieved June 23, 2025, from <https://isdr.org/networking/>
- Nicolosi Gelis, M. M., Cocherio, J., Donadelli, J., & Gómez, N. (2020a). Exploring the use of nuclear alterations, motility and ecological guilds in epipelagic diatoms as biomonitoring tools for water quality improvement in urban impacted lowland streams. *Ecological Indicators*, 110, 105951. <https://doi.org/10.1016/j.ecolind.2019.105951>
- Nicolosi Gelis, M. M., Cocherio, J., Donadelli, J., & Gómez, N. (2020b). Exploring the use of nuclear alterations, motility and ecological guilds in epipelagic diatoms as biomonitoring tools for water quality improvement in urban impacted lowland streams. *Ecological Indicators*, 110, 105951. <https://doi.org/10.1016/j.ecolind.2019.105951>
- Nicolosi Gelis, M. M., & Sathicq, M. B. (2020a). *diathor: Calculate Ecological Information and Diatom Based Indices* (p. 0.1.5) [Dataset]. <https://doi.org/10.32614/CRAN.package.diathor>
- Nicolosi Gelis, M. M., & Sathicq, M. B. (2020b). *diathor: Calculate Ecological Information and Diatom Based Indices* (p. 0.1.5) [Dataset]. <https://doi.org/10.32614/CRAN.package.diathor>
- Nicolosi Gelis, M. M., & Sathicq, M. B. (2020c). *diathor: Calculate Ecological Information and Diatom Based Indices* (p. 0.1.5) [Dataset]. <https://doi.org/10.32614/CRAN.package.diathor>
- Nicolosi Gelis, M. M., Sathicq, M. B., Jupke, J., & Cocherio, J. (2021). *DiaThor: R package for computing diatom metrics and biotic indices*. <https://doi.org/10.1101/2021.03.31.437895>
- Nicolosi Gelis, M. M., Sathicq, M. B., Jupke, J., & Cocherio, J. (2022a). *DiaThor: R package for computing diatom metrics and biotic indices*. *Ecological Modelling*, 465, 109859. <https://doi.org/10.1016/j.ecolmodel.2021.109859>
- Nicolosi Gelis, M. M., Sathicq, M. B., Jupke, J., & Cocherio, J. (2022b). *DiaThor: R package for computing diatom metrics and biotic indices*. *Ecological Modelling*, 465, 109859. <https://doi.org/10.1016/j.ecolmodel.2021.109859>
- Nicolosi Gelis, M. M., Sathicq, M. B., Jupke, J., & Cocherio, J. (2022c). *DiaThor: R package for computing diatom metrics and biotic indices*. *Ecological Modelling*, 465, 109859. <https://doi.org/10.1016/j.ecolmodel.2021.109859>
- Nicolosi Gelis, M.M., Sathicq, M.B., Jupke, J., & Cocherio, J. (2021). - Google Search. (n.d.). Retrieved May 14, 2025, from https://www.google.com/search?q=Nicolosi+Gelis%2C+M.M.%2C+Sathicq%2C+M.B.%2C+Jupke%2C+J.%2C+%26+Cochero%2C+J.+%282021%29.&sc=esv=1bbf12dcd0a2c&xsrf=AHTn8zqQu13Ce-butu2ALkWW5-NSOo9Emg%3A1747231072124&ei=YKEkaM6kBSCHvr0PzL6ngAs&ved=0ahUKEwjO1JP7jqONAxWQg68BHUzfCbAQ4dUDCBA&uact=5&oq=Nicolosi+Gelis%2C+M.M.%2C+Sathicq%2C+M.B.%2C+Jupke%2C+J.%2C+%26+Cochero%2C+J.+%282021%29.&gs_l=EGxnd3Mtd2l6LXNlcnAiRU5pY29sb3NpIEdlbGZlLCBNLk0uLCBTYXRoaWNxLCBNLkluLCBKdXBzSwgSi4sICyGQ29jaGVvbywgSi4gKDIwMjEpLkgAUABYAHAAcACOAQCYAOCgAOCqAOC4AOPIAOD4AOL4AOGYAqCgAgCYAwCSBwCgBwCyBwC4BwA&scient=gws-wiz-serp
- Oeding, S., & Taffs, K. H. (2017). Developing a regional diatom index for assessment and monitoring of freshwater streams in sub-tropical Australia. *Ecological Indicators*, 80, 135–146. <https://doi.org/10.1016/j.ecolind.2017.05.009>

Overview of Biological and Hydromorphological Assessment Methods in the Danube River Basin. (n.d.).

Pandey, L. Kr., Ojha, K. Kr., Singh, P. K., Singh, C. S., Dwivedi, S., & Bergey, E. A. (2016a). Diatoms image database of India (DIDI): A research tool. *Environmental Technology & Innovation*, 5, 148–160. <https://doi.org/10.1016/j.eti.2016.02.001>

Pandey, L. Kr., Ojha, K. Kr., Singh, P. K., Singh, C. S., Dwivedi, S., & Bergey, E. A. (2016b). Diatoms image database of India (DIDI): A research tool. *Environmental Technology & Innovation*, 5, 148–160. <https://doi.org/10.1016/j.eti.2016.02.001>

Pareek, R., Singh, G. P., & Singh, R. (n.d.). *Some fresh water diatoms of Galta kund, Jaipur, India.*

(PDF) A Coded Check List and Ecological Indicator Values of Freshwater Diatoms From the Netherlands. (2025). *ResearchGate*. <https://doi.org/10.1007/BF02334251>

(PDF) *Comparison of diatom-based indices of water quality of mid-continent (USA) great rivers.* (n.d.). *ResearchGate*. Retrieved July 7, 2025, from https://www.researchgate.net/publication/286751874_Comparison_of_diatom-based_indices_of_water_quality_of_mid-continent_USA_great_rivers

(PDF) *DiatomNet: An automatic Diatom genus identification system through microscopic images and Deep Learning.* (2025, March 29). *ResearchGate*. <https://doi.org/10.1101/2025.02.10.635050>

(PDF) *Diatoms image database of India (DIDI): A research tool.* (n.d.). Retrieved June 23, 2025, from https://www.researchgate.net/publication/343448147_Diatoms_image_database_of_India_DIDI_A_research_tool?utm_source=chatgpt.com

(PDF) Successive accumulation of biotic assemblages at a fine spatial scale along glacier-fed waters. (2024). *ResearchGate*. <https://doi.org/10.1016/j.isci.2024.109476>

(PDF) *Using diatom life-forms and ecological guilds to assess organic pollution and trophic level in rivers: A case study of rivers in south-eastern France.* (n.d.). Retrieved May 14, 2025, from https://www.researchgate.net/publication/227310241_Using_diatom_life-forms_and_ecological_guilds_to_assess_organic_pollution_and_trophic_level_in_rivers_a_case_study_of_rivers_in_south-eastern_France

Posit. (n.d.-a). *Posit*. Retrieved May 16, 2025, from <https://www.posit.co/>

Posit. (n.d.-b). *Posit*. Retrieved May 16, 2025, from <https://www.posit.co/>

Pu, S., Zhang, F., Shu, Y., & Fu, W. (2023). Microscopic image recognition of diatoms based on deep learning. *Journal of Phycology*, 59(6), 1166–1178. <https://doi.org/10.1111/jpy.13390>

Ratheesh Kumar, R., Vineetha, G., & Monolisha, S. (2024). *Plankton taxonomy* (No. 41; Issue 41, pp. 155–174). ICAR- Central Marine Fisheries Research Institute. <http://eprints.cmfri.org.in/18443/>

Rimet, F. (n.d.). *Larras F., Bouchez A., Rimet F. & Montuelle B. 2012. Using Bioassays and Species Sensitivity Distributions to Assess Herbicide Toxicity towards Benthic Diatoms. PLoS ONE 7 (8).* <https://doi.org/10.1371/JOURNAL.PONE.0044458>

Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A., & Bouchez, A. (2016). R-Syst::diatom: An open-access and curated barcode database for diatoms and freshwater monitoring. *Database*, 2016, baw016. <https://doi.org/10.1093/database/baw016>

Search Diatom Image Database. (n.d.). Retrieved June 23, 2025, from <https://websites.rbge.org.uk/ADIAC/db/adiacdb.htm>

Search results | Taylor & Francis Online. (n.d.). Retrieved July 7, 2025, from <https://www.tandfonline.com/action/doSearch?AllField=%09SI%C3%A1de%C4%8Dek%27s+Saprobic+Index&SeriesKey=tinw19&startPage=&pageSize=10>

Selvaratnam, N. (2023). *How to Run Essential Analyses in R Studio: For First Year Undergraduate Students with No Background in Programming* (Eagan Journal of Contemporary Research, Ed.; First Edition). Deep Haven Counselling (Pvt) Ltd (10.61495). <https://doi.org/10.61495/ejcr.0001.23>

Sládeček, V., & Sládečková, A. (1998). Revision of polysaprobic indicators. *SIL Proceedings, 1922-2010*, 26(3), 1277–1280. <https://doi.org/10.1080/03680770.1995.11900928>

- Spaulding, S. A., Potapova, M. G., Bishop, I. W., Lee, S. S., Gasperak, T. S., Jovanoska, E., Furey, P. C., & Edlund, M. B. (2021a). *Diatoms.org*: Supporting taxonomists, connecting communities. *Diatom Research*, 36(4), 291–304. <https://doi.org/10.1080/0269249X.2021.2006790>
- Spaulding, S. A., Potapova, M. G., Bishop, I. W., Lee, S. S., Gasperak, T. S., Jovanoska, E., Furey, P. C., & Edlund, M. B. (2021b). *Diatoms.org*: Supporting taxonomists, connecting communities. *Diatom Research*, 36(4), 291–304. <https://doi.org/10.1080/0269249X.2021.2006790>
- Spaulding, S. A., Potapova, M. G., Bishop, I. W., Lee, S. S., Gasperak, T. S., Jovanoska, E., Furey, P. C., & Edlund, M. B. (2021c). *Diatoms.org*: Supporting taxonomists, connecting communities. *Diatom Research*, 36(4), 291–304. <https://doi.org/10.1080/0269249X.2021.2006790>
- Spaulding, S. A., Potapova, M. G., Bishop, I. W., Lee, S. S., Gasperak, T. S., Jovanoska, E., Furey, P. C., & Edlund, M. B. (2021d). *Diatoms.org*: Supporting taxonomists, connecting communities. *Diatom Research*, 36(4), 291–304. <https://doi.org/10.1080/0269249X.2021.2006790>
- Spaulding, S. A., Potapova, Marina G., Bishop, Ian W., Lee, Sylvia S., Gasperak, Tim S., Jovanoska, Elena, Furey, Paula C., & and Edlund, M. B. (2021). *Diatoms.org*: Supporting taxonomists, connecting communities. *Diatom Research*, 36(4), 291–304. <https://doi.org/10.1080/0269249X.2021.2006790>
- Srivastava, P., Verma, J., Grover, S., & Sardar, A. (2016). *ON THE IMPORTANCE OF DIATOMS AS ECOLOGICAL INDICATORS IN RIVER ECOSYSTEMS: A REVIEW*.
- Subrahmanyam, R. (1946). A systematic account of the marine plankton diatoms of the Madras coast. *Proceedings / Indian Academy of Sciences*, 24(4), 85–197. <https://doi.org/10.1007/BF03049673>
- Szczepocka, E., & Szulc, B. (2009). The use of benthic diatoms in estimating water quality of variously polluted rivers. *Oceanological and Hydrobiological Studies*, 38(1), 17–26. <https://doi.org/10.2478/v10009-009-0012-x>
- Tapolczai, K., Chonova, T., Fidlerová, D., Makovinská, J., Mora, D., Weigand, A., & Zimmermann, J. (2024). Molecular metrics to monitor ecological status of large rivers: Implementation of diatom DNA metabarcoding in the Joint Danube Survey 4. *Ecological Indicators*, 160, 111883. <https://doi.org/10.1016/j.ecolind.2024.111883>
- Taxonomic richness in the diatom flora of Himalayan streams (Garhwal, India)*. (n.d.). Retrieved June 23, 2025, from https://www.jstage.jst.go.jp/article/diatom1985/20/0/20_123/article/-char/ja/
- Teittinen, A., Taka, M., Ruth, O., & Soininen, J. (2015a). Variation in stream diatom communities in relation to water quality and catchment variables in a boreal, urbanized region. *Science of The Total Environment*, 530–531, 279–289. <https://doi.org/10.1016/j.scitotenv.2015.05.101>
- Teittinen, A., Taka, M., Ruth, O., & Soininen, J. (2015b). Variation in stream diatom communities in relation to water quality and catchment variables in a boreal, urbanized region. *Science of The Total Environment*, 530–531, 279–289. <https://doi.org/10.1016/j.scitotenv.2015.05.101>
- The measures of saprobity: With 6 figures and 5 tables in the text: SIL Proceedings, 1922-2010: Vol 17, No 1*. (n.d.). Retrieved July 7, 2025, from <https://www.tandfonline.com/doi/abs/10.1080/03680770.1968.11895887>
- Tree of Science with Scopus: A Shiny Application | Issues in Science and Technology Librarianship*. (n.d.). Retrieved May 14, 2025, from <https://journals.library.ualberta.ca/istl/index.php/istl/article/view/2698>
- Trophic and saprobic diatom-based indices are not indicating the same ecological status in a shallow lake, Türkiye | Request PDF. (n.d.). *ResearchGate*. <https://doi.org/10.1007/s42974-024-00199-1>
- “UDE DIATOMS in the Wild 2024”: A new image dataset of freshwater diatoms for training deep learning models | Request PDF. (2024). *ResearchGate*. <https://doi.org/10.1093/gigascience/giae087>
- Valentin, V., Frédéric, R., Isabelle, D., Olivier, M., Yorick, R., & Agnès, B. (2019). Assessing pollution of aquatic environments with diatoms’ DNA metabarcoding: Experience and developments from France Water Framework Directive networks. *Metabarcoding and Metagenomics*, 3, e39646. <https://doi.org/10.3897/mbmg.3.39646>
- Venkataraman, G. (1939). A systematic account of some South Indian diatoms. *Proceedings / Indian Academy of Sciences*, 10(6), 293–368. <https://doi.org/10.1007/BF03039988>

- Villar, P., Casado, J., Fernández, D., Sánchez, P., & Tabik, S. (2025). *DiatomNet: An automatic Diatom genus identification system through microscopic images and Deep Learning* (p. 2025.02.10.635050). bioRxiv. <https://doi.org/10.1101/2025.02.10.635050>
- Wickham, H. (2021). *Mastering Shiny*. O'Reilly Media, Inc.
- Wood, R. J., Mitrovic, S. M., Lim, R. P., Warne, M. St. J., Dunlop, J., & Kefford, B. J. (2019a). Benthic diatoms as indicators of herbicide toxicity in rivers – A new SPEcies At Risk (SPEARherbicides) index. *Ecological Indicators*, 99, 203–213. <https://doi.org/10.1016/j.ecolind.2018.12.035>
- Wood, R. J., Mitrovic, S. M., Lim, R. P., Warne, M. St. J., Dunlop, J., & Kefford, B. J. (2019b). Benthic diatoms as indicators of herbicide toxicity in rivers – A new SPEcies At Risk (SPEARherbicides) index. *Ecological Indicators*, 99, 203–213. <https://doi.org/10.1016/j.ecolind.2018.12.035>
- Yu, W., Xiang, Q., Hu, Y., Du, Y., Kang, X., Zheng, D., Shi, H., Xu, Q., Li, Z., Niu, Y., Liu, C., & Zhao, J. (2022). An improved automated diatom detection method based on YOLOv5 framework and its preliminary study for taxonomy recognition in the forensic diatom test. *Frontiers in Microbiology*, 13. <https://doi.org/10.3389/fmicb.2022.963059>
- भारतीय प्राणी सर्वेक्षण की आधिकारिक वेबसाइट, पर्यावरण, वन और जलवायु परिवर्तन मंत्रालय, भारत सरकार. (n.d.). Retrieved June 23, 2025, from <https://zsi.gov.in/impo-link/en?v=7>
- Ács, É., Szabo, K., Tóth, B., & Kiss, K. T. (2004). Investigation of benthic algal communities, especially diatoms of some Hungarian streams in connection with reference conditions of the water framework directives. *Acta Botanica Hungarica*, 46, 255–278. <https://doi.org/10.1556/ABOT.46.2004.3-4.1>
- Bąk, M., Witkowski, A., LANGE-BERTALOT, H., & Dadal, A. (2004). Ecology of the Szczecin Lagoon diatom flora with reference to the utility of diatom indices in assessing water quality. *Diatom : The Japanese Journal of Diatomology*, 20, 23–31. https://doi.org/10.11464/DIATOM1985.20.0_23
- Battegazzore, M., Gallo, L., Lucadamo, L., & Morisi, A. (2004). *Quality of the main watercourses in the Pollino National Park (Apennine Mts , S Italy) on the basis of the Diatom benthic communities*. <https://www.semanticscholar.org/paper/0e9911f23f8e6450ca4a0c1894b7c7648c60bba6>
- Bere, T. (2016a). Are diatom-based biotic indices developed in eutrophic, organically enriched waters reliable monitoring metrics in clean waters? *Ecological Indicators*, 62, 312–316. <https://doi.org/10.1016/J.ECOLIND.2015.11.008>
- Bere, T. (2016b). Challenges of diatom-based biological monitoring and assessment of streams in developing countries. *Environmental Science and Pollution Research*, 23, 5477–5486. <https://doi.org/10.1007/s11356-015-5790-y>
- Bere, T., Mangadze, T., & Mwedzi, T. (2014). The application and testing of diatom-based indices of stream water quality in Chinhoyi Town, Zimbabwe. *Water SA*, 40, 503–512. <https://doi.org/10.4314/WSA.V40I3.14>
- Bere, T., Phiri, C., Kadye, W. T., & Utete, B. (2013). Benthic diatom assemblages in mountain streams: Community structure in relation to environmental and human pressures. *African Journal of Ecology*, 51, 625–634. <https://doi.org/10.1111/AJE.12078>
- Bere, T., & Tundisi, J. (2011a). Applicability of borrowed diatom-based water quality assessment indices in streams around São Carlos-SP, Brazil. *Hydrobiologia*, 673, 179–192. <https://doi.org/10.1007/s10750-011-0772-7>
- Bere, T., & Tundisi, J. (2011b). Influence of land-use patterns on benthic diatom communities and water quality in the tropical Monjolinho hydrological basin, São Carlos-SP, Brazil. *Water SA*, 37, 93–102. <https://doi.org/10.4314/WSA.V37I1.64112>
- Besse-Lototskaya, A., Verdonchot, P., & Sinkeldam, J. (2006). Uncertainty in Diatom Assessment: Sampling, Identification and Counting Variation. *Hydrobiologia*, 566, 247–260. <https://doi.org/10.1007/s10750-006-0092-5>
- Bytyçi, P., Ymeri, P., Czikkely, M., Fetoshi, O., Shala-Abazi, A., Ismaili, M., Ramshaj, Q., & Millaku, F. (2019). The Application of Benthic Diatoms in Water Quality Assessment in Lepenci River Basin, Kosovo. *Journal of Ecological Engineering*, 20, 43–57. <https://doi.org/10.12911/22998993/113409>

- Cappelletti, C., Ciutti, F., Crippa, A., Mancini, L., Beltrami, M., Pierdominici, E., & Dell'uomo, A. (2005). *Diatomee come indicatori della qualità biologica dei corsi d'acqua: EPI-D ed altri metodi europei a confronto: Il caso studio del fiume Tevere*. <https://www.semanticscholar.org/paper/e308382194a18ca4628d267a38d2513326b77034>
- Ciutti, F., Cappelletti, C., & Corradini, F. (2004). *Applicazione dell'indice EPI-D a un corso d'acqua delle Alpi (torrente Fersina): Osservazioni sulla metodica di determinazione delle abbondanze relative*. <https://www.semanticscholar.org/paper/16bf461ea95a5ada6f5480c61e1bdca2ba8d6a38>
- Dalu, T., Bere, T., & Froneman, P. (2016). Assessment of water quality based on diatom indices in a small temperate river system, Kowie River, South Africa. *Water SA*, 42, 183–193. <https://doi.org/10.4314/WSA.V42I2.02>
- Delgado, C., Pardo, I., & García, L. (2010). A multimetric diatom index to assess the ecological status of coastal Galician rivers (NW Spain). *Hydrobiologia*, 644, 371–384. <https://doi.org/10.1007/s10750-010-0206-y>
- Duleba, M., Földi, A., Micsinai, A., Mohr, A., Sipos, R., Szabó, G., Trábert, Z., Bíró, T., Buczkó, K., Várbiro, G., & Ács, É. (2021). Applicability of diatom metabarcoding in the ecological status assessment of Hungarian lotic and soda pan habitats. *ARPHA Conference Abstracts*, null, null. <https://doi.org/10.3897/ACA.4.E64887>
- Eloranta, P., & Soininen, J. (2002). Ecological status of some Finnish rivers evaluated using benthic diatom communities. *Journal of Applied Phycology*, 14, 1–7. <https://doi.org/10.1023/A:1015275723489>
- Gelis, M. M. N., Cocherio, J., Donadelli, J., & Gómez, N. (2020). Exploring the use of nuclear alterations, motility and ecological guilds in epipellic diatoms as biomonitoring tools for water quality improvement in urban impacted lowland streams. *Ecological Indicators*, null, null. <https://doi.org/10.1016/j.ecolind.2019.105951>
- Gelis, M. N., Sathicq, M. B., Jupke, J., & Cocherio, J. (2021). DiaThor: R package for computing diatom metrics and biotic indices. *bioRxiv*, null, null. <https://doi.org/10.1101/2021.03.31.437895>
- Gonçalves, V., Raposeiro, P., & Costa, A. C. (2008). Benthic diatoms and macroinvertebrates in the assessment of the ecological status of Azorean streams. *Limnetica*, null, null. <https://doi.org/10.23818/limn.27.25>
- Harding, W., Archibald, C., & Taylor, J. C. (2005). The relevance of diatoms for water quality assessment in South Africa: A position paper. *Water SA*, 31, 41–46. <https://doi.org/10.4314/WSA.V31I1.5119>
- Kalyoncu, H., & Şerbetci, B. (2013). Applicability of Diatom-Based Water Quality Assessment Indices in Dari Stream, Isparta- Turkey. *World Academy of Science, Engineering and Technology, International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering*, 7, 386–394.
- Kryk, A., Bąk, M., Kaniak, A., & Adamczyk, M. (2022). Is it possible to optimise the labour and time intensity of diatom analyses for determination of the Polish Diatom Indices (IO, IOJ)? *Environmental Monitoring and Assessment*, 195, null. <https://doi.org/10.1007/s10661-022-10676-7>
- Kwandrans, J., Eloranta, P., Kawecka, B., & Wojtan, K. (1998). Use of benthic diatom communities to evaluate water quality in rivers of southern Poland. *Journal of Applied Phycology*, 10, 193–201. <https://doi.org/10.1023/A:1008087114256>
- Lecointe, C., Coste, M., & Prygiel, J. (1993). “Omnidia”: Software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, 269–270, 509–513. <https://doi.org/10.1007/BF00028048>
- Prygiel, J., & Coste, M. (1993). The assessment of water quality in the Artois-Picardie water basin (France) by the use of diatom indices. *Hydrobiologia*, 269–270, 343–349. <https://doi.org/10.1007/BF00028033>
- Qu, X., Zhou, Y., Zhao, R., Bentsen, C. N., Yin, X., & Zhang, Y. (2014). Are diatom-based indices from Europe suitable for river health assessment in China? A case study from Taizi River, northeastern China. *British Journal of Environment and Climate Change*, 4, 95–114. <https://doi.org/10.9734/BJECC/2014/8577>
- Rakowska, B., & Szczepocka, E. (2011). Demonstration of the Bzura River restoration using diatom indices. *Biologia*, 66, 411–417. <https://doi.org/10.2478/s11756-011-0032-3>
- Rey, P., Taylor, J. C., Laas, A., Rensburg, L., & Vosloo, A. (2004). Determining the possible application value of diatoms as indicators of general water quality: A comparison with SASS 5. *Water SA*, 30, 325–332. <https://doi.org/10.4314/WSA.V30I3.5080>

- Rimet, F., Ciutti, F., Cappelletti, C., & Ector, L. (2005). *Ruolo delle diatomee nell'applicazione della direttiva europea quadro sulle acque*. <https://www.semanticscholar.org/paper/3c1fcd59c0137de0e6dbbd9ee1244dd10546209>
- Salinas-Camarillo, V. H., Carmona-Jiménez, J., & Lobo, E. A. (2020). Development of the Diatom Ecological Quality Index (DEQI) for peri-urban mountain streams in the Basin of Mexico. *Environmental Science and Pollution Research*, 28, 14555–14575. <https://doi.org/10.1007/s11356-020-11604-3>
- Soeprbowati, T., Purnaweni, H., Jumari, J., & Sari, K. (2021). The Relationship of Water Quality to Epipellic Diatom Assemblages in Cebong Lake, Dieng Indonesia. *Polish Journal of Environmental Studies*, null, null. <https://doi.org/10.15244/pjoes/137084>
- Solak, C., Peszek, Ł., Yılmaz, E., Ergül, H., Kayal, M., Ekmekçi, F., Várбірó, G., Yüce, A., Canlı, O., Binici, M. S., & Ács, É. (2020). Use of Diatoms in Monitoring the Sakarya River Basin, Turkey. *Water*, null, null. <https://doi.org/10.3390/w12030703>
- Szczepocka, E., Szulc, B., Szulc, K., Rakowska, B., & Żelazna-Wieczorek, J. (2014). Diatom indices in the biological assessment of the water quality based on the example of a small lowland river. *Oceanological and Hydrobiological Studies*, 43, 265–273. <https://doi.org/10.2478/S13545-014-0141-Z>
- Szczepocka, E., & Żelazna-Wieczorek, J. (2018). Diatom biomonitoring – scientific foundations, commonly discussed issues and frequently made errors. *Oceanological and Hydrobiological Studies*, 47, 313–325. <https://doi.org/10.1515/OHS-2018-0030>
- Tan, X., Liu, Y., Burford, M., & Zhang, Q. (2020). The performance of diatom indices in assessing temporal changes in water quality in a large lowland river ecosystem. *River Research and Applications*, 37, 423–432. <https://doi.org/10.1002/rra.3754>
- Tan, X., Sheldon, F., Bunn, S., & Zhang, Q. (2013). Using diatom indices for water quality assessment in a subtropical river, China. *Environmental Science and Pollution Research*, 20, 4164–4175. <https://doi.org/10.1007/s11356-012-1343-9>
- Taylor, J. C., Harding, W., Archibald, C., & Rensburg, L. V. (2007). Diatoms as indicators of water quality in the Jukskei-Crocodile river system in 1956 and 1957, a re-analysis of diatom count data generated by BJ Chohnoky. *Water SA*, 31, 237–246. <https://doi.org/10.4314/WSA.V31I2.5176>
- Taylor, J. C., Prygiel, J., Vosloo, A., Rey, P. D. L., & Rensburg, L. (2007). Can diatom-based pollution indices be used for biomonitoring in South Africa? A case study of the Crocodile West and Marico water management area. *Hydrobiologia*, 592, 455–464. <https://doi.org/10.1007/s10750-007-0788-1>
- Taylor, J., Vuuren, M., & Pieterse, A. (2009). The application and testing of diatom-based indices in the Vaal and Wilge Rivers, South Africa. *Water SA*, 33, 51–59. <https://doi.org/10.4314/WSA.V33I1.47871>
- Wu, J.-T., & Kow, L.-T. (2002). Applicability of a generic index for diatom assemblages to monitor pollution in the tropical River Tsanwun, Taiwan. *Journal of Applied Phycology*, 14, 63–69. <https://doi.org/10.1023/A:1015277013102>

GLOSSARY

1. *Abundance* — The total number of individuals of a particular species found in a given sample or habitat. In *DiaThor*, abundance data form the foundation of index calculations.
2. *Absolute Abundance* — The raw, unscaled number of diatom valves or cells counted in a sample before converting to percentages or relative values.
3. *Acidophilous Diatoms* — Species that prefer acidic water conditions (pH below 6.5), often found in peat bogs and forested wetlands.
4. *Aerophilous Species* — Diatoms that thrive in moist terrestrial or semi-aquatic environments such as rocks, mosses, and stream edges exposed to air.
5. *Algorithm* — A defined sequence of steps or instructions used to perform calculations or automate analyses, such as index computation in *DiaThor*.
6. *Alkalibiontic Diatoms* — Species confined to highly alkaline waters (pH > 8.5).
7. *Alkaliphilous Diatoms* — Species that prefer slightly to moderately alkaline conditions (pH 7.5–8.5).
8. *Alpha Diversity* — A measure of species diversity within a single site or sample, commonly represented by species richness or Shannon index.
9. *Alt Text (Alternative Text)* — Descriptive text used in digital figures and plots to ensure accessibility and context in publications.
10. *Anoxic Conditions* — Environments lacking dissolved oxygen; often reflected by dominance of tolerant diatom species.
11. *API (Application Programming Interface)* — A framework enabling software components to communicate, such as linking RStudio scripts with web databases.
12. *Aquatic Biofilm* — A slimy layer formed by microorganisms (including diatoms) on submerged surfaces, serving as a habitat and food source.
13. *AWS (Amazon Web Services)* — A cloud computing platform often used for data storage and processing in MLOps pipelines.
14. *Azoic Zone* — Layers of sediment or water bodies devoid of living organisms due to extreme conditions.
15. *Bar Plot* — A graphical display of categorical data using rectangular bars; often used in *DiaThor* for index comparison.
16. *Baseline Data* — Initial set of observations used as a reference for detecting future changes in environmental monitoring.
17. *Batch Processing* — Automated analysis of multiple datasets simultaneously, implemented in advanced *DiaThor* CSV plotters.
18. *Benthic Diatoms* — Diatom species that live attached to sediments, rocks, or plants at the bottom of water bodies.
19. *Bioaccumulation* — The gradual buildup of substances like metals or toxins within an organism over time.
20. *Bioconcentration Factor (BCF)* — A ratio expressing how much of a chemical accumulates in an organism compared to its environment.
21. *Biofilm* — A cohesive microbial community adhering to a surface, where diatoms play a key role in structure and nutrient cycling.
22. *Bioindicator* — An organism that provides information on the quality or change of an environment based on its presence, absence, or abundance.
23. *Biomonitoring* — The systematic use of biological responses to assess environmental conditions or contamination levels.

24. *Biotic Index* — A numerical expression of ecological quality derived from the composition and abundance of species sensitive to pollution.
25. *Biovolume* — The estimated three-dimensional volume of diatom cells, representing biomass and ecological adaptation.
26. *Bootstrap Sampling* — A statistical resampling technique in R to estimate the variability or confidence intervals of an index.
27. *Brackish Water* — Water with salinity between freshwater and seawater, typically found in estuaries and mangroves.
28. *Cache Memory* — Temporary data storage in RStudio or computer memory that improves performance during iterative code execution.
29. *Calibration* — The process of adjusting instruments or models to ensure accurate environmental or computational measurements.
30. *Carbon Fixation* — The process by which diatoms and other photosynthetic organisms convert CO₂ into organic compounds.
31. *Centric Diatoms* — Diatoms with radial symmetry, typically planktonic and dominant in marine or open-water systems.
32. *Checksum Error* — A data validation error occurring when file integrity is compromised; sometimes triggered by corrupted CSV uploads.
33. *Chloroplast Morphology* — The number, shape, and arrangement of chloroplasts within diatoms, used in functional trait analysis.
34. *Chorology* — The study of the spatial distribution of species, applicable to diatom biogeography.
35. *Clean Code* — A software engineering practice emphasizing readable, maintainable, and efficient programming.
36. *Cluster Analysis* — A statistical method that groups sites or samples based on similarity in species composition or environmental parameters.
37. *Code Chunk* — A block of executable code in R Markdown or Quarto documents used for reproducible analyses.
38. *Coding Environment* — The integrated setup (software, packages, libraries) used for programming — e.g., RStudio for R users.
39. *Colonial Diatoms* — Species that form chains, ribbons, or colonies instead of living as solitary cells.
40. *Comma-Separated Values (CSV)* — A simple file format for storing tabular data. Required input format for DiaThor analysis.
41. *Community Composition* — The list and relative abundance of species present in an ecosystem or dataset.
42. *Conductivity* — A measure of the water's ability to conduct electricity, linked to ion concentration and pollution.
43. *Console* — The main command line interface in RStudio where users execute code and view outputs.
44. *Containerization* — The practice of packaging code and dependencies using tools like Docker for portability across systems.
45. *Correlation Coefficient* — A numerical measure of the relationship between two variables, often used to assess index-environment relationships.
46. *CRAN (Comprehensive R Archive Network)* — The primary repository for R packages, including DiaThor.
47. *Cron Job* — A scheduled automated task used in data pipelines or MLOps environments.
48. *Data Cleaning* — The process of identifying and correcting errors or inconsistencies in datasets before analysis.

49. *Data Dictionary* — A table describing every variable and column in a dataset, detailing its meaning, type, and expected format.
50. *Data Frame* — The fundamental R object structure for tabular data, similar to an Excel spreadsheet.
51. *Data Pipeline* — A series of automated steps that extract, transform, and load data for analysis or visualization.
52. *Data Visualization* — The graphical representation of data patterns or relationships through plots, graphs, or charts.
53. *Database Mismatch* — A common error in DiaThor when species names in input data do not match reference databases.
54. *Decibel Plot* — A logarithmic plot used in environmental acoustics; metaphorically similar to scaling ecological gradients.
55. *Descy Index (IDG)* — A diatom-based metric sensitive to organic pollution, developed by Jean-Pierre Descy.
56. *DiaThor_all()* — The master function in DiaThor that performs all available analyses in one automated workflow.
57. *DiaThor Output Folder* — The directory automatically created by DiaThor containing CSV files, logs, and visualization plots.
58. *Diatom* — A unicellular alga with a silica shell (frustule), forming one of the most ecologically significant groups in aquatic ecosystems.
59. *Diatom Database (Diat.Barcode)* — A global reference dataset linking diatom taxa to ecological traits, tolerance, and biogeographic data.
60. *Diversity Index* — A statistical measure combining species richness and evenness, e.g., Shannon or Simpson index.
61. *Docker* — A platform that isolates software applications within lightweight containers for reproducibility.
62. *Dominance* — The degree to which a few species account for most of the individuals in a community.
63. *Ecosystem Functioning* — The collective biological, geochemical, and physical processes that sustain life in an ecosystem.
64. *Ecological Guild* — A group of diatom species sharing similar traits and resource-use strategies.
65. *Ecological Integrity* — The wholeness and functioning of an ecosystem's structure and processes, often inferred from index scores.
66. *Ecotone* — The transitional zone between two ecological communities, often supporting high biodiversity.
67. *Endemic Species* — Species restricted to a specific geographic area or habitat.
68. *Entropy* — In ecology, a measure of diversity; in computing, a measure of randomness in data.
69. *Error Log* — A diagnostic file automatically generated by DiaThor to report issues during execution.
70. *Epilithic Diatoms* — Species that grow on rock surfaces in aquatic habitats.
71. *Epiphytic Diatoms* — Species that attach to aquatic plants or macrophytes.
72. *Episammic Diatoms* — Species living attached to sand grains.
73. *Eutrophication* — The process by which a water body becomes enriched with nutrients, leading to excessive plant and algal growth.
74. *FastAPI* — A modern web framework for building APIs, commonly used in data science deployment.
75. *Field Sampling* — The process of collecting biological or environmental samples from natural habitats.
76. *Filtering Step* — A data preprocessing stage in R where unwanted or erroneous records are removed.
77. *Frustule* — The siliceous, glass-like cell wall of a diatom, often intricately patterned and species-specific.

78. *Functional Trait* — A measurable biological characteristic influencing an organism's ecological role (e.g., size, motility, guild).
79. *GIS (Geographic Information System)* — Software used to map and analyze spatial data linked to sampling sites.
80. *Guild Composition* — The percentage representation of diatom functional groups in a community.
81. *Habitat Stability* — The persistence and consistency of environmental conditions that influence diatom community structure.
82. *Heatmap* — A data visualization technique using color gradients to represent matrix-like values such as ecological indices.
83. *Heterotrophy* — The ecological mode of obtaining energy by consuming organic matter rather than producing it via photosynthesis.
84. *Hydrology* — The scientific study of water movement, distribution, and quality across ecosystems.
85. *Index Coverage* — The range and completeness of indices included within a dataset or software like DiaThor.
86. *Indicator Value (V)* — A weighting factor representing the ecological reliability of a species within an index formula.
87. *Installation Path* — The directory location on a computer where software like DiaThor or R packages are stored.
88. *Intercalibration* — The process of aligning different ecological indices or monitoring methods for consistency across studies.
89. *Interpolation* — Estimating unknown data points within the range of known observations.
90. *IPS (Indice de Polluo-Sensibilité Spécifique)* — A key diatom index indicating organic pollution sensitivity; higher values mean cleaner water.
91. *JSON (JavaScript Object Notation)* — A lightweight format for data interchange, sometimes used in R APIs and DiaThor outputs.
92. *Key Species* — Diatom taxa with high influence on community structure or index calculation due to abundance or sensitivity.
93. *Legend* — The explanatory component of a plot or map describing symbols, colors, or variables.
94. *Library (in R)* — A collection of functions packaged together to extend R's capabilities.
95. *Linear Regression* — A statistical method for modeling the relationship between two continuous variables.
96. *Lithophilic Species* — Diatoms preferring rocky substrates for attachment.
97. *Macroinvertebrates* — Aquatic organisms visible to the naked eye; often used alongside diatoms in biomonitoring.
98. *Metadata* — Data describing other data (e.g., date, site, analyst); essential for reproducibility.
99. *Migration* — The movement of organisms from one habitat to another, affecting community composition.
100. *Model Validation* — The process of verifying that a computational model accurately represents observed ecological patterns.
101. *Motile Diatoms* — Species capable of movement through mucilage secretion; indicators of sediment instability.
102. *Multivariate Analysis* — A suite of statistical tools (e.g., PCA, NMDS) for analyzing complex ecological data.
103. *NA (Not Available)* — A missing data symbol in R that must be handled properly to avoid calculation errors.
104. *Normalization* — Scaling data to a standard range for comparison, common in trait and index visualizations.
105. *Nutrient Enrichment* — Increase in nitrogen or phosphorus levels in water bodies, often leading to eutrophication.

- 106. *Observation Unit* — The smallest unit of analysis, typically a sample site or date.
- 107. *Open Source* — Software whose source code is freely available for modification and redistribution, like R and DiaThor.
- 108. *Outlier* — A data point that deviates significantly from other observations, potentially indicating measurement or entry error.
- 109. *Package Dependency* — Another package required for a software library to run properly in R.
- 110. *Parameter* — A variable that defines system behavior or model settings.
- 111. *PCA (Principal Component Analysis)* — A technique for reducing dataset dimensionality while preserving variance, used for ecological ordination.
- 112. *Phytoplankton* — Microscopic algae floating in water; diatoms are a dominant group among them.
- 113. *Pipeline* — A structured sequence of code steps automating data processing.
- 114. *Pixel Resolution* — The quality and clarity of an image or plot measured in pixels per inch (PPI).
- 115. *Plot Aesthetics* — Visual attributes (color, size, font) that define the appearance of a graph.
- 116. *Pollution Gradient* — A continuum from unpolluted to heavily polluted conditions represented by index variation.
- 117. *Precision* — The consistency or repeatability of measurements.
- 118. *Reactivity (in Shiny)* — A feature in R Shiny that updates outputs dynamically when input values change.
- 119. *Relative Abundance* — The proportion of a species relative to the total individuals in a sample.
- 120. *Reproducibility* — The ability to recreate identical analytical results using the same data and code.
- 121. *Resolution* — The level of detail in digital plots or raster images.
- 122. *Sample ID* — A unique alphanumeric identifier for each sample or site.
- 123. *Sampling Bias* — Systematic error introduced by uneven or non-random sampling methods.
- 124. *Saprobity* — Degree of organic pollution in a water body, classified into oligosaprobic, mesosaprobic, or polysaprobic categories.
- 125. *Script* — A text file containing R code to automate repetitive tasks.
- 126. *Sensitivity Value (S)* — A numerical score representing how sensitive a species is to pollution.
- 127. *Shannon Index* — A diversity measure combining species richness and evenness.
- 128. *Snowflake / Airflow* — Cloud-based tools for managing large-scale data pipelines in MLOps.
- 129. *Species Evenness* — The uniformity of species distribution in a sample.
- 130. *Species Richness* — The number of unique species present in a dataset.
- 131. *SPEAR (Species at Risk)* — A diatom index measuring sensitivity to herbicides or pesticides.
- 132. *SQL (Structured Query Language)* — Used to manage and query databases integrated with environmental data.
- 133. *Standard Deviation* — A statistical measure of variability in data.
- 134. *Station Code* — A label identifying a monitoring site, often tied to geographic coordinates.
- 135. *STL (Seasonal-Trend Decomposition)* — An R function used for analysing seasonal patterns in time series data.
- 136. *Taxonomic Hierarchy* — The classification of species into genus, family, order, etc.

- 137. TDI (Trophic Diatom Index) — A widely used metric for nutrient enrichment assessment.*
- 138. Time Series — Sequential data collected over time, useful for tracking environmental change.*
- 139. Trait-Based Analysis — Examining ecological function using traits rather than taxonomy.*
- 140. UTF-8 Encoding — A universal file encoding ensuring special characters in CSVs load correctly in R.*
- 141. Van Dam Classification — A trait-based ecological classification of diatoms by preference for salinity, oxygen, pH, and trophic state.*
- 142. Variable — A measurable attribute in a dataset.*
- 143. Vector Search — A machine-learning retrieval method used in AI, conceptually similar to ecological similarity analysis.*
- 144. Vegan Package — A popular R library for community ecology and diversity analysis.*
- 145. Version Control — Systematic tracking of code changes using tools like Git or GitHub.*
- 146. Visualization — The practice of representing data graphically to reveal patterns and relationships.*
- 147. Water Quality Index (WQI) — A composite score summarizing multiple environmental parameters.*
- 148. Weighted Mean — The average of values where each value is multiplied by a weight, as in index formulas.*
- 149. Workflow — The structured progression of analytical steps from raw data to final output.*
- 150. Zero-Inflated Data — Datasets with excessive zero values, common in ecological abundance tables.*

MESSAGE FROM THE AUTHOR

I was born in Prayagraj, a city where faith meets flow, where rivers embrace and part again, much like the stories of our own lives. I grew up as a small-town girl with quiet dreams and an ever-curious mind. Books were my closest companions, and ambition often whispered to me through the pages I turned. I had charted my own course or so I thought but destiny, with its subtle grace, had other plans.

Life, I have come to realize, never rises in a straight line. The curve of growth is uneven, filled with troughs, pauses, and sudden ascents. Each fall, each turn, carried a lesson I was too young then to understand. There were times I resisted, for the road ahead did not match the one I had imagined. Yet, it is often the path we resist most that reveals our truest strength.

Research was never my chosen calling. It entered my life quietly, almost uninvited, like a gentle rain that stays longer than you expect, changing the scent of the soil forever. And somewhere along the way, I began to see purpose in the unexpected. Perhaps this too was divine design, God's subtle way of reminding me that we are not always the authors of our beginnings, but we may still shape the grace of our becoming.

My father taught me the essence that lies at the heart of every sacred text, humanity. His life and words have been a compass, guiding me through the moments when reason faltered and only faith remained.

Today, I walk with faith, not in certainty, but in surrender. For I have learned that destiny does not always answer our prayers; sometimes, it fulfills them in disguise. And in those disguises, I have found not what I wanted, but what I needed; a journey shaped by resilience, faith, and the quiet hand of the Creator who writes our story between the lines we never planned to read. There comes a moment in every one's life when our work becomes a reflection of the self.

This book, "Indexing Diatom with DiaThor in R Studio," is born not just of data and code, but of faith, endurance, and the quiet strength that life teaches through its trials.

Each error corrected, each index generated, became a step closer to understanding not only the ecology of rivers, but the deeper currents of resilience within the human spirit. The diatoms, microscopic yet magnificent, whispered lessons of patience, harmony, and balance. They taught me that greatness often resides in gentleness, and that persistence, not perfection, defines true strength. To my Supervisor, Dr. Jyoti Verma, Assistant Professor in Zoology, C.M.P. Degree College, words fall short of what your mentorship has meant to me. You have been my light through confusion, my anchor in turbulence, and my guiding star when I lost my way.

Like an angel in academic disguise, you believed in me when I doubted myself, guided me with patience, and inspired me with grace.

This book exists because you once said, "You can do it, Arpita."

I come from a lineage where words breathe wisdom and faith flows through generations.

My grandfather (Late Mr. S.N Lal Srivastava), man of words, passed down the sacred inheritance of knowledge, not as an obligation, but as a way of life. His pen was his prayer, and perhaps, through his genes, that calling has found me too.

My grandmother (Late Mrs. Triveni Srivastava), whose love was quiet and unconditional, shaped the rhythm of my heart. She taught me that strength is not loud, it smiles softly, forgives quietly, and keeps believing even when the world does not.

My father (Mr. S.P Srivastava), instilled in me the philosophy that true wisdom is the union of intellect and integrity. From him I learned that purpose must always be higher than pride, and that faith, steady and unshaken, is life's most powerful equation.

My mother (Mrs. Dolly Srivastava), the embodiment of kindness, taught me compassion even when the world grows cruel. Her strength wears gentleness as its armor, and her voice remains my calm through every storm.

My Uncle (Mr. B.P Srivastava), a man of rare dignity and discipline, and living proof that brilliance and humility can coexist. From him I learned that true success is measured not by position, but by poise, by the grace with which one handles both power and people.

My elder brother (Late Mr. Rijul Srivastava), left this world in 2014, but never my side. He was my first teacher, my greatest believer, my forever guide. His laughter still echoes in my determination; his belief still fuels my courage. This book, every word of it, carries his soul, for it was he who taught me to turn pain into purpose and silence into strength.

My husband (Mr. Subir Verma), who is a Senior Data Scientist, is the reason the language of code feels like poetry to me. In him, I found both a companion in life and a collaborator in thought, his analytical brilliance often became the bridge between my ideas and their realization. He gave structure to my work. Life, in its infinite wisdom, brings every person into our path for a purpose. I pause to reflect upon the many hearts that have silently walked beside me, some with words, others with wisdom, and a few with quiet distance that still taught me strength. To all my extended family members, friends and well-wishers who have stood by me in their own silent ways, I remain deeply grateful.

And now, as I stand at the threshold of motherhood, waiting to welcome my little one into this world, I know this book is not just mine.

It belongs also to the child I carry, who has already been a silent companion through my writing days and sleepless nights.

I want that maybe, when you one day read this book, know that you were the heartbeat behind every page, a promise of new beginnings, the purest inspiration for perseverance.

As Rudyard Kipling once wrote in 'If':

"If you can meet with Triumph and Disaster

And treat those two impostors just the same..."

Those words have lived with me throughout this journey, reminding me that real success is not in applause, but in endurance; not in the result, but in the resolve.

This book is not just a scientific guide, it is a tribute to every person, known or unseen, who believed, encouraged, or challenged me along the way.

It is my offering to science, to family, to faith and to the little soul who will soon call me “mother.”

May it remind every reader that behind every dataset lies a dream, behind every code a story, and behind every researcher, a heart that dares to hope.

With gratitude and hope.

Arpita Srivastava